UTD 2016 Spring

CS6301.5U1 Advanced Computational Methods for Data Science

Assignment 5 – Subset Selection and Shrinkage

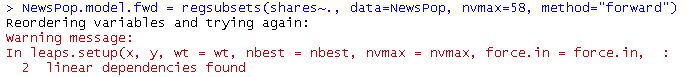
Name: Yu Zhang, Mingyue Sun

NetID: yxz141631, mxs151730

# Subset Selection

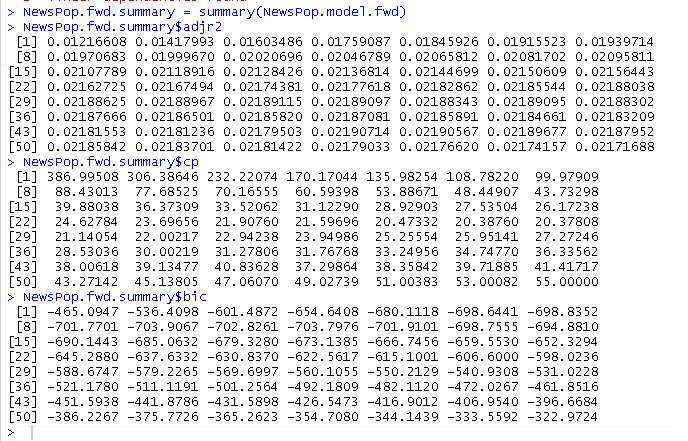
## Forward Subset Selection

Using “regsubset” function with “method = forward” attribute to do the backward subset selection.

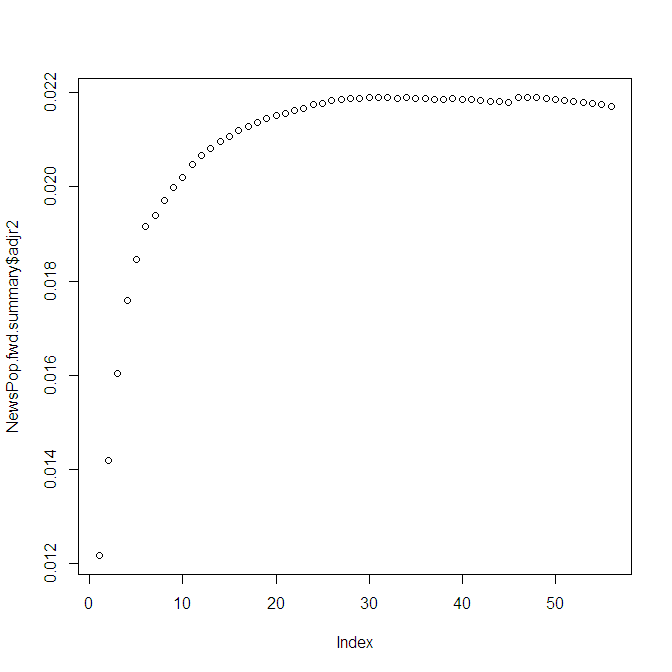


The warning shown above indicates that there are two columns that are linear dependent on each other. So we will only see 56 attributes in the results.

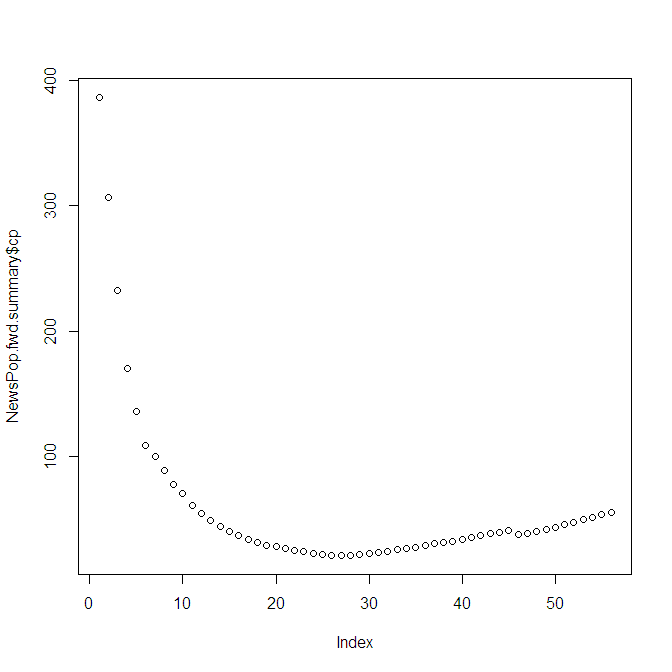
**1.1.1** Subset Selection Based on Adjusted R2, Cp, BIC



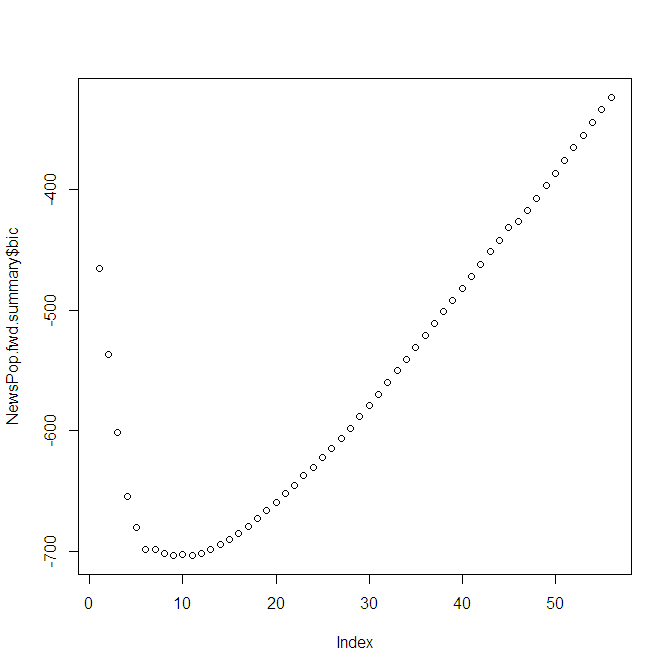
To view the details more clearly, we can plot the coefficients on schemes.



**Shown above is the plot for Adjusted R2**

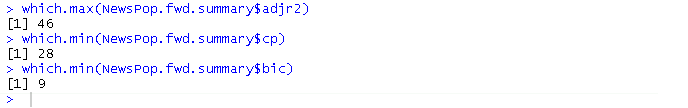


**Shown above is the plot for Cp**

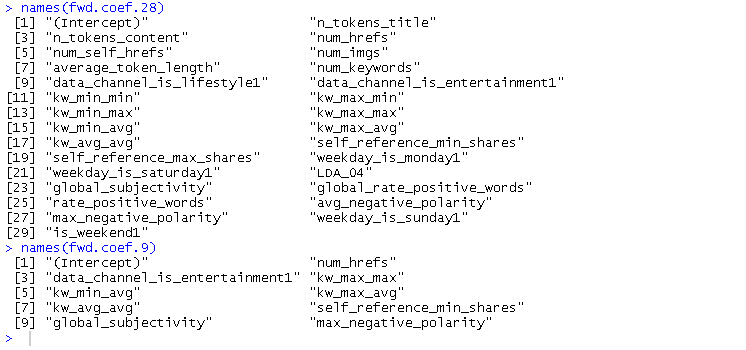
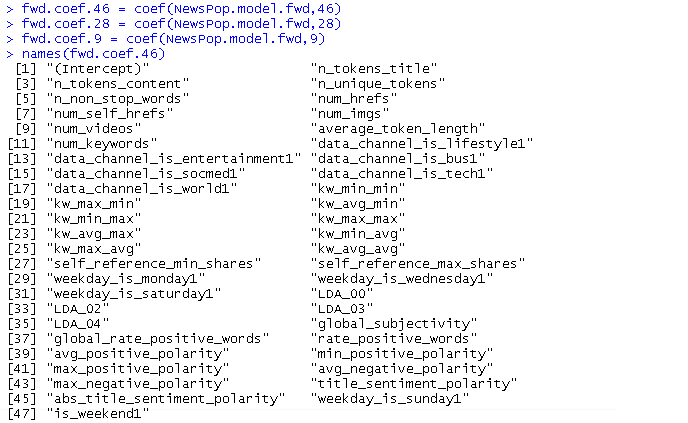


**Shown above is the plot for BIC**

To choose the best subsets we will base on maximum R2, minimum Cp and minimum BIC separately.

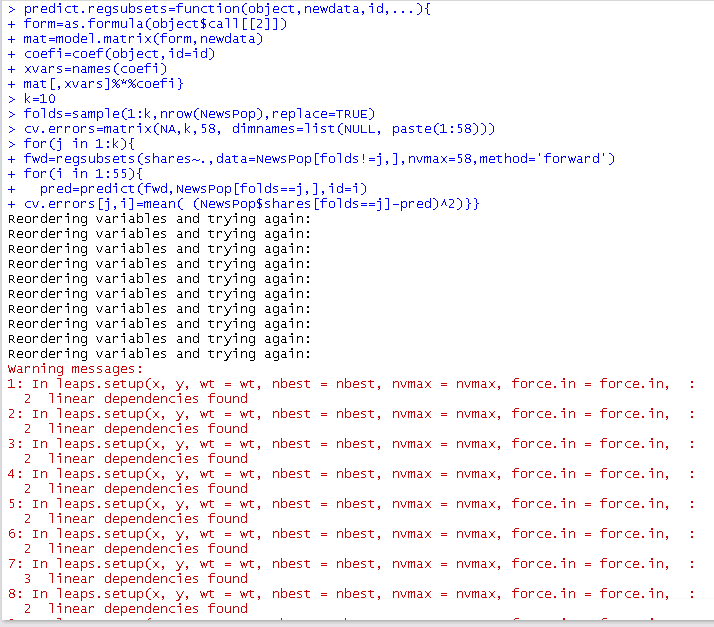


Here are the subsets chosen basing on the R maximum R2, minimum Cp and minimum BIC from forward subset selection.

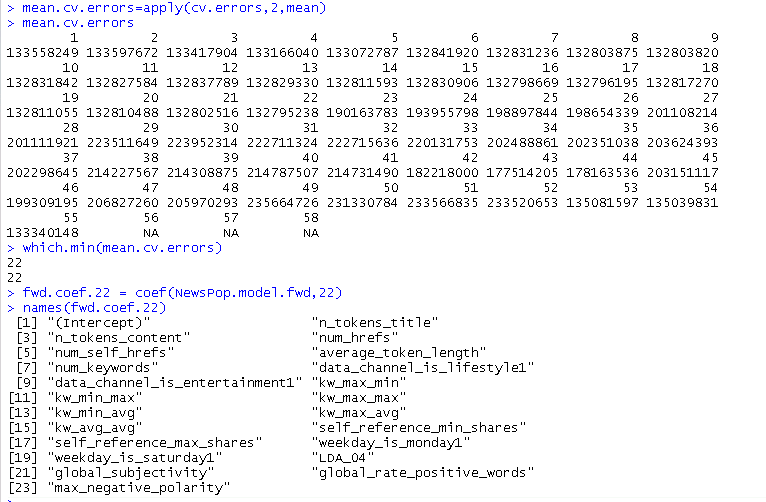


### 1.1.2 Subset Selection Based on Cross Validation

In the following part we will use 10 folds cross validation to choose the best subset number with forward subset selection method.



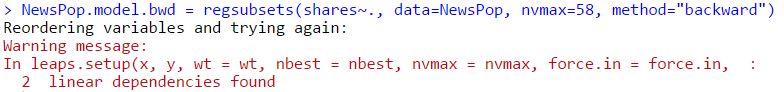
From the warnings reported we can clearly indicate that there are 2 columns linearly correlated with other column. And we will choose the number of subsets that has the lowest mean square error in cross validation result.



From the result we can indicate that the best number of subset is 22. And the column names are shown as above.

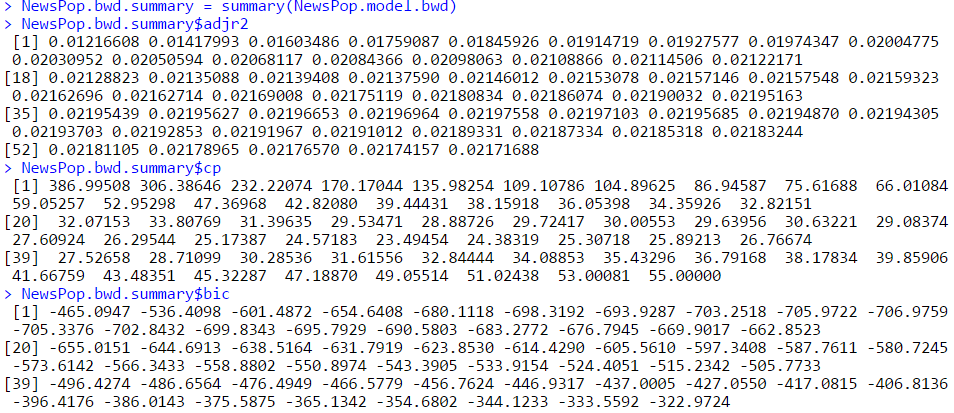
## 1.2Backward Subset Selection

Using “regsubset” function with “method = backward” attribute to do the backward subset selection.

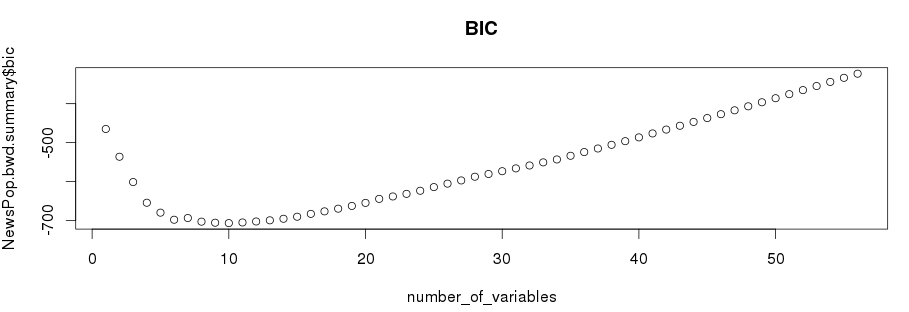
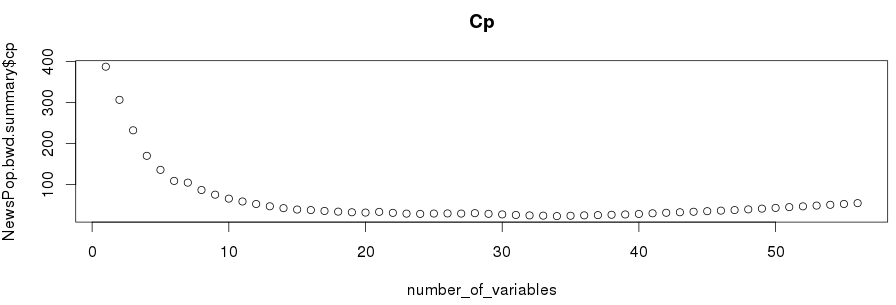
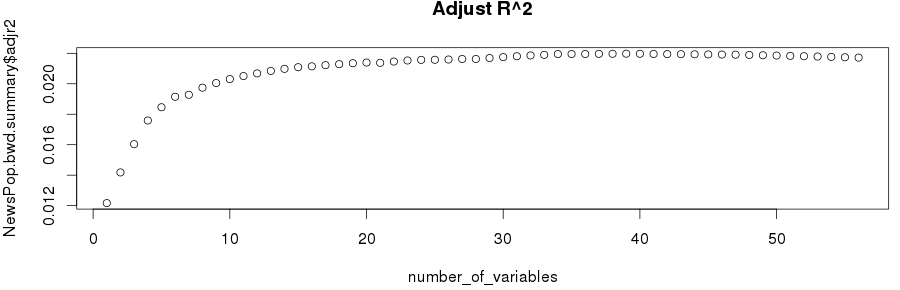


A warning is shown that there are two columns that are linear dependent on each other. So we will only see 56 attributes in the results.

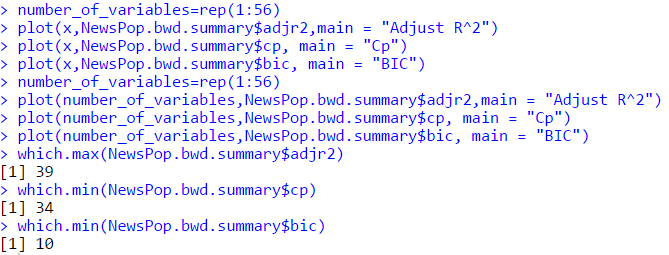
### 1.2.1 Subset Selection Basing on Adjust R2, Cp, BIC



To view the details more clearly, we can plot the coefficients on schemes.



### To choose the best subsets we will base on maximum R2, minimum Cp and minimum BIC separately.



Here are the subsets chosen basing on the R maximum R2, minimum Cp and minimum BIC from backward subset selection.

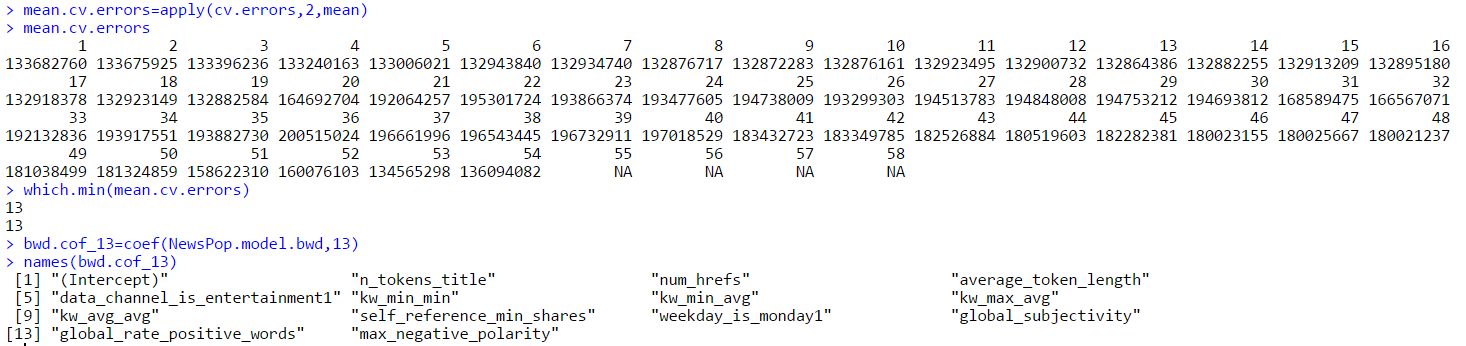


### 1.2.2 Subset Selection Basing on Cross Validation

In the following part we will use 10 folds cross validation to choose the best subset number with backward subset selection method.



From the warnings reported we can clearly indicate that there are 2 columns linearly correlated with other column. And we will choose the number of subsets that has the lowest mean square error in cross validation result.



From the result we can indicate that the best number of subset is 13. And the column names are shown as above.

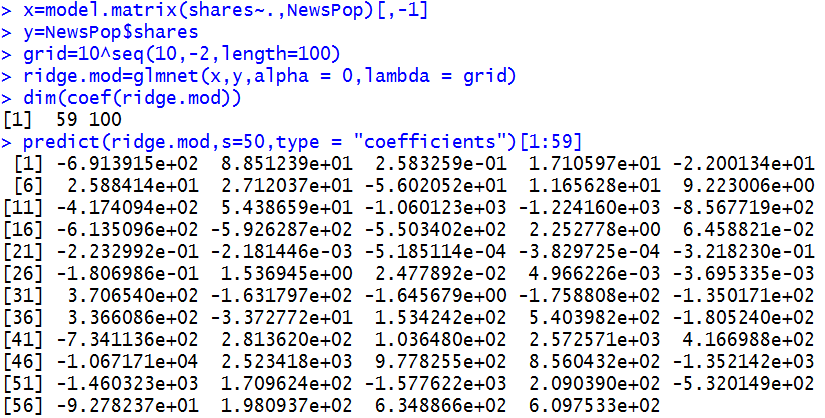
# Shrinkage

## RIDGE

Ridge Regression is a technique for analyzing multiple regression data that suffer from multicollinearity. When multicollinearity occurs, least squares estimates are unbiased, but their variances are large so they may be far from the true value. By adding a degree of bias to the regression estimates, ridge regression reduces the standard errors. It is hoped that the net effect will be to give estimates that are more reliable. Another biased regression technique, principal components regression, is also available in NCSS. Ridge regression is the more popular of the two methods.

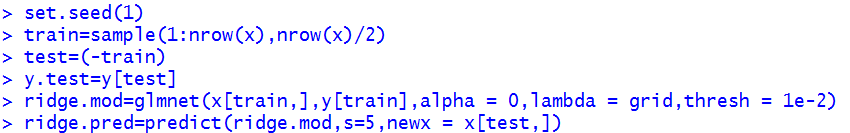
### Normal Ridge Shrinkage

In the following we will use RIDGE regression method provided by “glmnet” package to regress the data in linear model using lambdas ranging from 1010 to 1012 which has 100 numbers in total. And predict the coefficients using lambda equals 50.

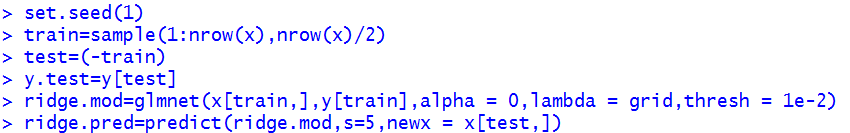


### Lambda Selection Basing on Cross Validation

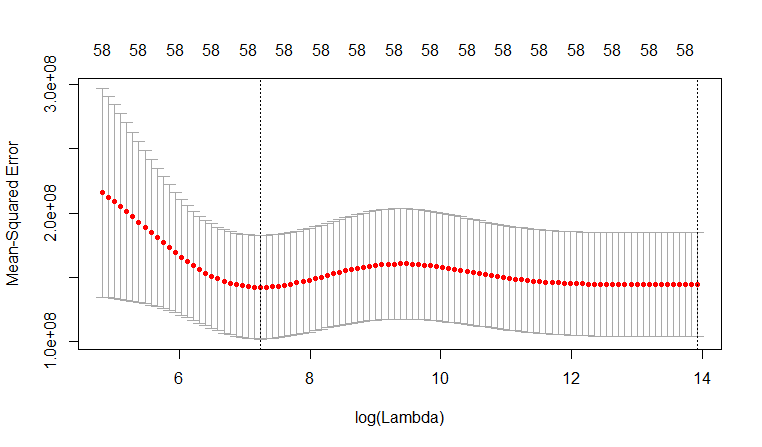
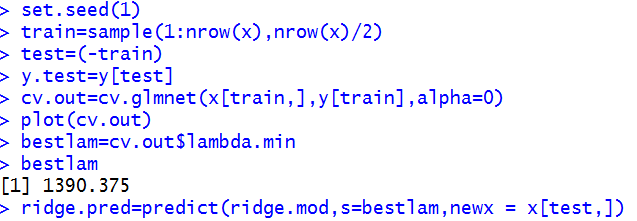
Separating the dataset into training (50%) and testing (50%) using normal RIDGE regression method.



The MSE calculation is as following.



Now we will use the cross validation to find the best lambda value and use it to calculate the new MSE.



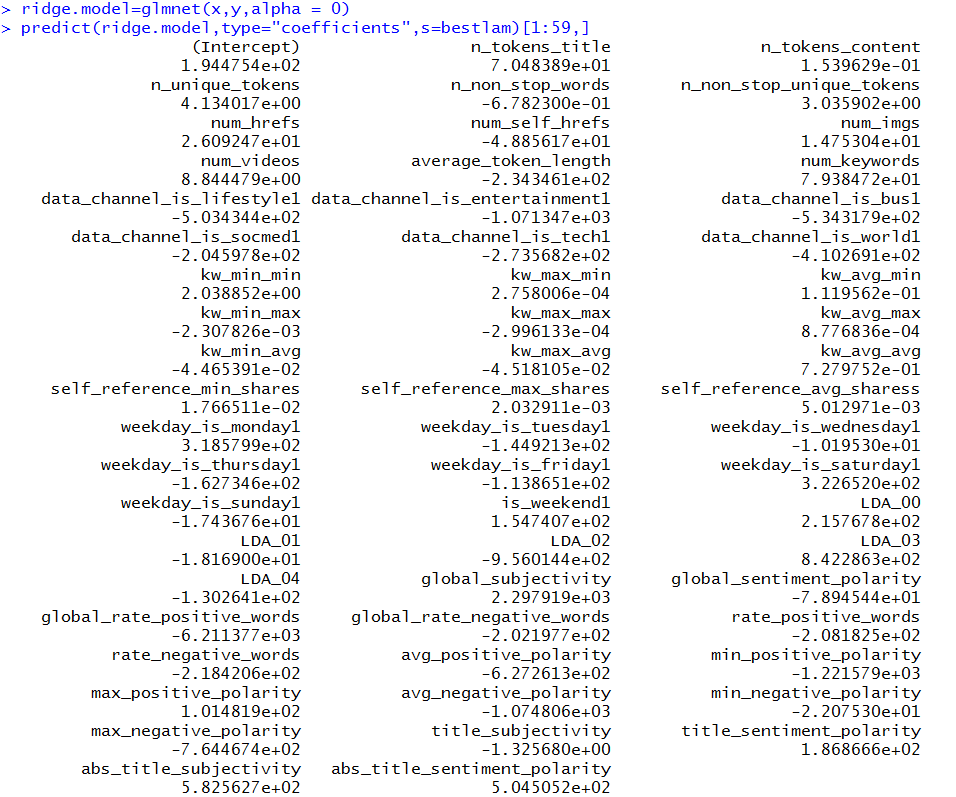
Now we can find the best lambda equals 1390.375. We will apply this in the following MSE calculation.



We can clearly indicate that the MSE has reduced by using cross validation method to find the best lambda.

### Final Coefficients

Using the best lambda, we can find the final coefficients that can build the best model.

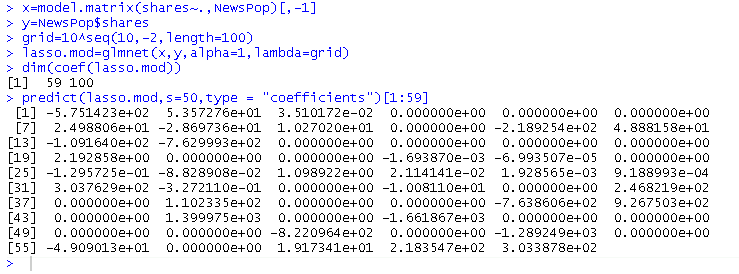


* 1. LASSO

In statistics and machine learning, **lasso (least absolute shrinkage and selection operator)** (also Lasso or LASSO) is a regression analysismethod that performs both variable selection and regularization in order to enhance the prediction accuracy and interpretability of the statistical model it produces. It was introduced by Robert Tibshirani in 1996 based on Leo Breiman’s Nonnegative Garrote.Lasso was originally formulated for least squares models and this simple case reveals a substantial amount about the behavior of the estimator, including its relationship to ridge regression and best subset selection and the connections between lasso coefficient estimates and so-called soft thresholding. It also reveals that (like standard linear regression) the coefficient estimates need not be unique if covariates are collinear.

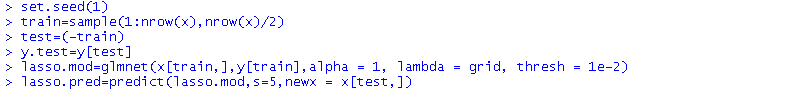
### Normal LASSO Shrinkage

In the following we will use LASSO regression method provided by “glmnet” package to regress the data in linear model using lambdas ranging from 1010 to 1012 which has 100 numbers in total. And predict the coefficients using lambda equals 50.



### Lambda Selection Basing on Cross Validation

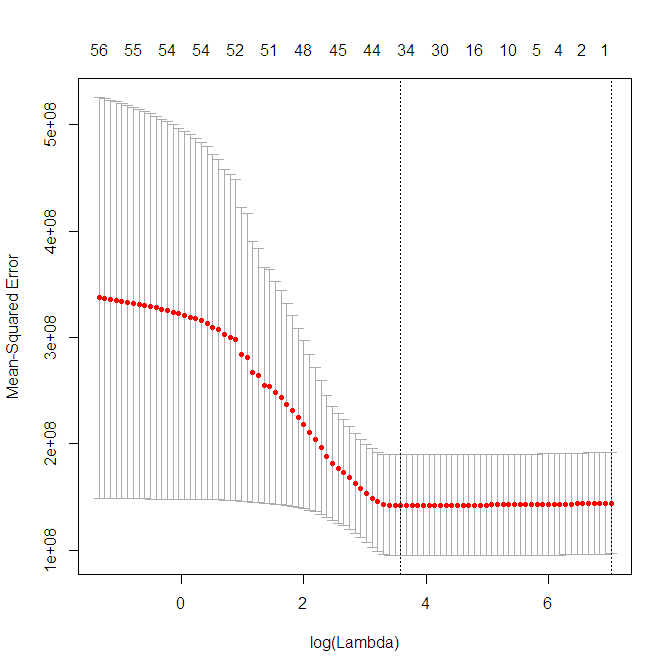
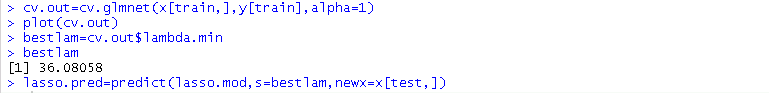
Separating the dataset into training (50%) and testing (50%) using normal RIDGE regression method.



The MSE calculation is as following.



Now we will use the cross validation to find the best lambda value and use it to calculate the new MSE.



We can clearly indicate that the MSE has reduced by using cross validation method to find the best lambda.



### Final Coefficients

Using the best lambda, we can find the final coefficients that can build the best model.

