**Question 2 HW4**

The question asks to apply Logistic Regression and LDA on weekly dataset on ISLR Package.

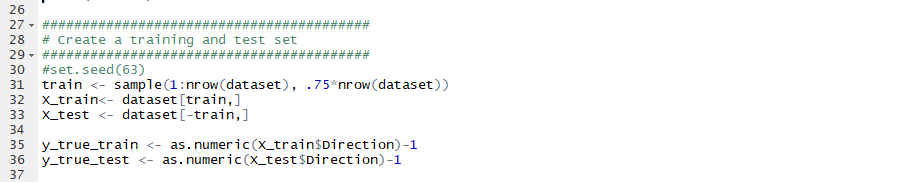
**Step 1 Loading the dataset**



**Step 2 Splitting the dataset**

the seed() function is set to a random variable to get the same values when executed again. The data is spitted into 75:25 ratio so that 75% of the data is used for training and remaining 25% is used for testing

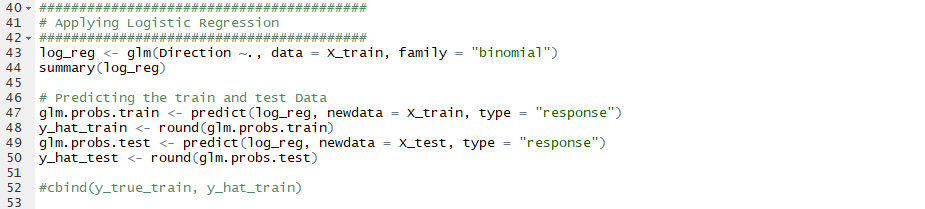
the categorical variable Direction, which is the dependent variable, is numerically encoded into 0 and 1 as shown below.

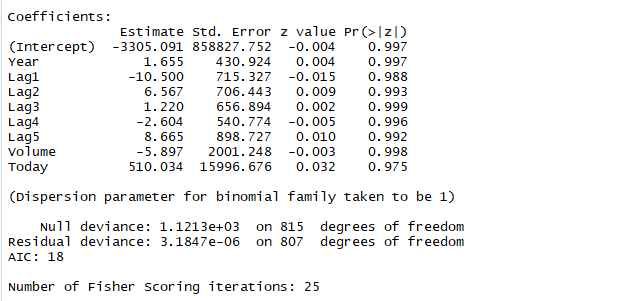


**Step 3 Applying Logistic regression**

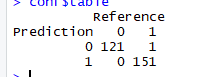
I applied the **Logistic regression on** the dataset using glm() function using glmnet library and **family** is set to binomial as there are only two classes in the dependent variable, summary is shown below by using summary() function.

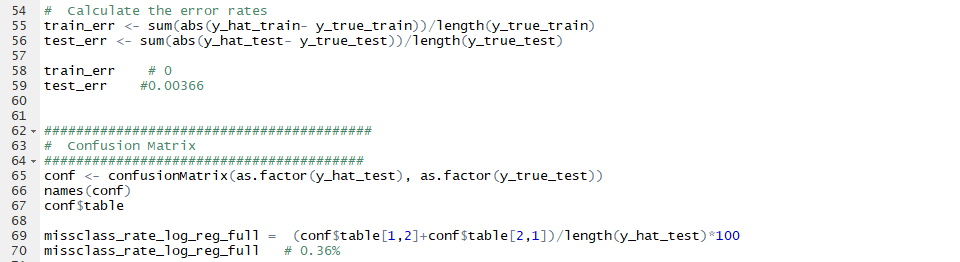
The results of test data are predicted by fittin the data to the test data.

****

* each variable appears to be significant in predicting the dependent variable, with each having same Z-score almost.

**Step 4 2C calculating the errors & Confusion Matrix**

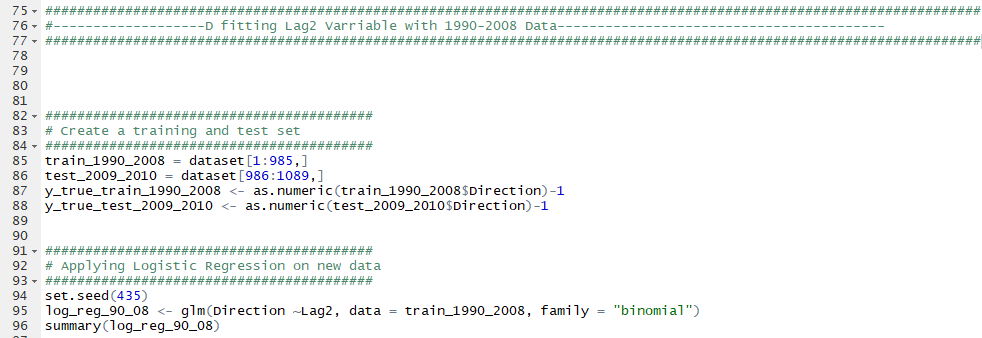
On calculating the train error and test error based on the predictions made by the model, it appeared that the train error is **0**  test\_error = **0.036%.** And there is only one instance of miss classification in the test set.and the model appered To be predicting accurately.**the misclassifiaction rate is 0.36%** .



**Step 5 (2)D Applying Logistic regression on 1990-2008 DATA**

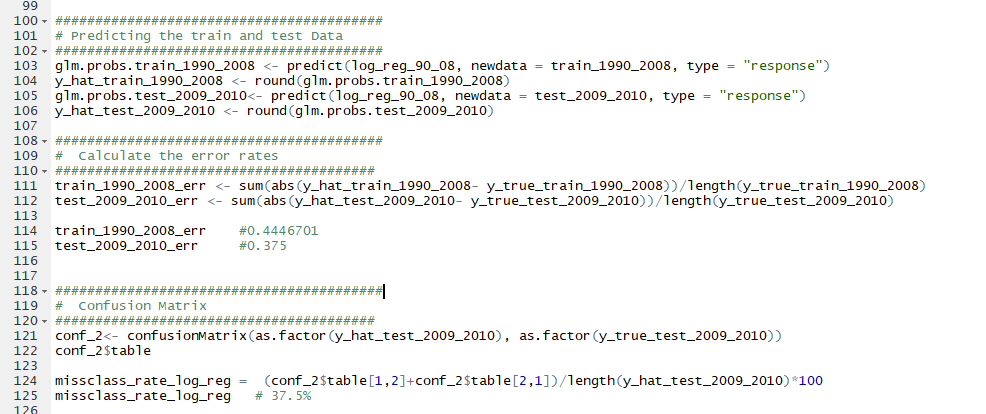
The question asks to take 90-08 data and also Lag1 as independent variable and predict the direction, I created the new trainset and test set with the above features.and aslo made the categoriacal variable numeric form.

I applied the **Logistic regression on** the dataset using glm() function using glmnet library and **family** is set to binomial as there are only two classes in the dependent variable, summary is shown below by using summary() function.



**Step 6 Predicting and Confusion Matrix**

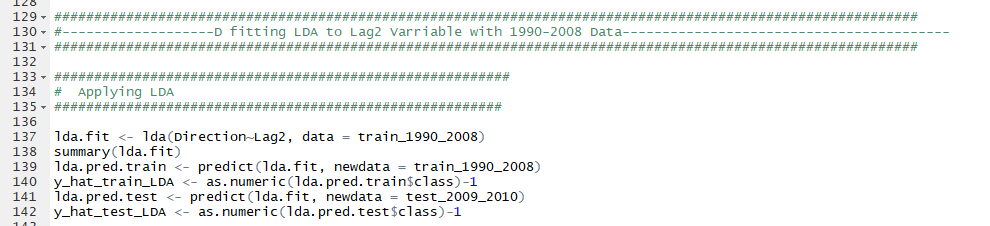
On calculating the train error and test error based on the predictions made by the model, it appeared that the train error is **0.444**  test\_error = **0.375.** And there is only one instance of miss classification in the test set.and the model appered To be predicting accurately.**the misclassifiaction rate is 0.37.5%** slightly higher than the Logistic reggression model.



**Step 7 Applying LDA on the SAME data**

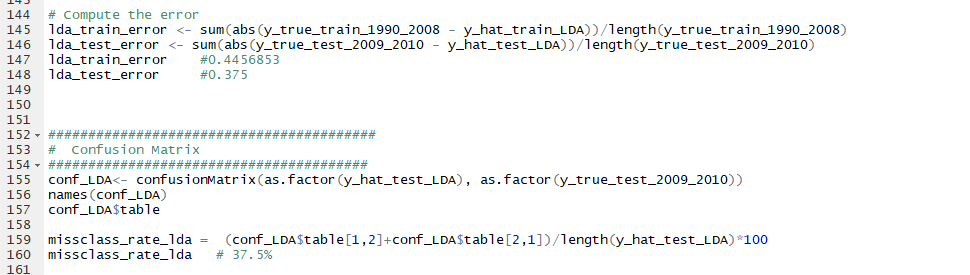
I applied the **LDA** onthe dataset using lda(), with LAG2 as the only varibale and Direction as dependent variable on data from between years 1990-2008 for training and remaining data for testing.

The results of test data are predicted by fittin the data to the test data, by making them as numerical factors of 0 and 1.

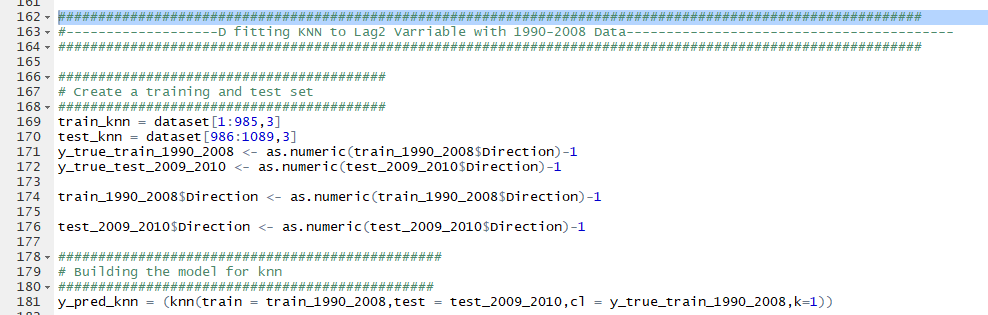


**Step 8 Predicting and Confusion Matrix(LDA)**

On calculating the train error and test error based on the predictions made by the model, it appeared that the train error is **0.445**  test\_error = **0.375.** And there is only one instance of miss classification in the test set.and the model appered To be predicting accurately.**the misclassifiaction rate is 0.37.5%**

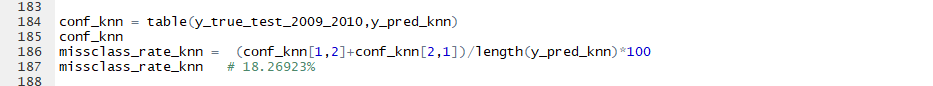
* it appers that the results are almost same as of logistic regression withg same test and train errors

**Step 9 Applying KNN on the SAME data**

Here I fitted the KNN model by taking Lag2~Direction by splitting the dataset first, and also converting the categorical variavle into numerical factors.

**Step 10 Predicting and Confusion Matrix(LDA)**the miss classification rate appears to be 18.26% less when compaed to the above two models

* It is appered that knn performs better than the above two models
* LDA and logistic regression produce same results and are simliar but they have there differences in them.



**Step 11 (2 A) Analysing the data**

* From the pair wise scatterplots of the variables it is found that
* The 5 Lag variables are not correlated to each other.
* The volume appers to be incresing along with the years from 1990-2010
* The 5 lag variables appears to be same all over the years with no much difference.

