This question should be answered using the Weekly data set, which is part of the ISLR2 package. This data is similar in nature to the Smarket data from this chapter’s lab, except that it contains 1, 089 weekly returns for 21 years, from the beginning of 1990 to the end of 2010.

1. Produce some numerical and graphical summaries of the Weekly data. Do there appear to be any patterns?

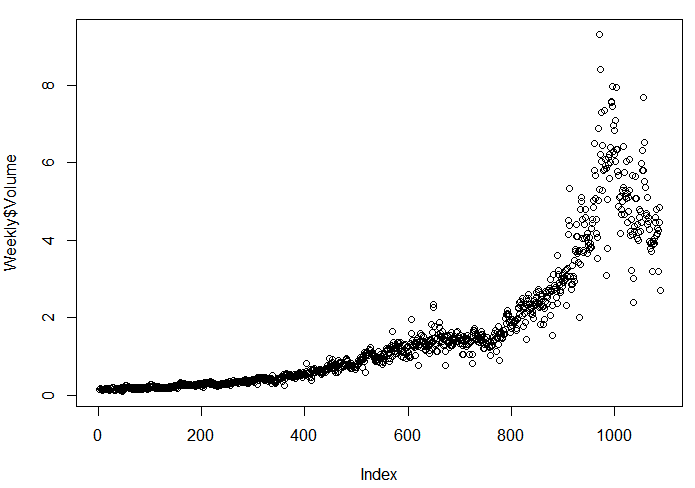
Because most of the data is about time lag, the summary information is very similar.

The lowest weekly return week is in 2008 W40, and it because the Financial crisis in 2008

一張含有 文字, 計分板, 螢幕擷取畫面 的圖片

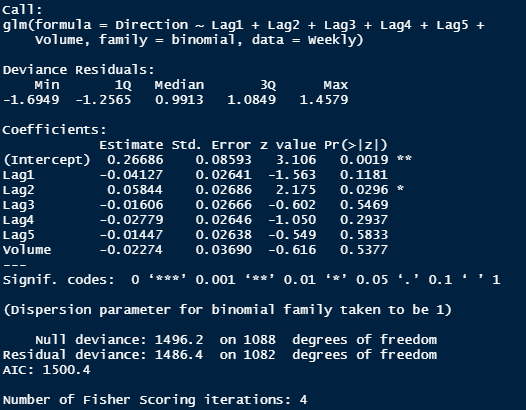
自動產生的描述

The trading volume increasing in the data, in 1990 there was only 0.15 billions every day, but in 2010 there was 4 billions every day.



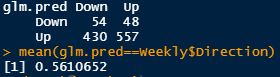
Overall, the weeks goes up(605) is more than the weeks goes down(484).

1. Use the full data set to perform a logistic regression with Direction as the response and the five lag variables plus Volume as predictors. Use the summary function to print the results. Do any of the predictors appear to be statistically significant? If so, which ones?

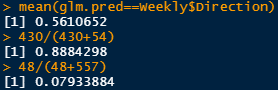


Under the significant level at 0.05, the predictor “Lag2” is statistically significant.

1. Compute the confusion matrix and overall fraction of correct predictions. Explain what the confusion matrix is telling you about the types of mistakes made by logistic regression.



The overall accuracy is only about 56% which means it not a perfect model.

We have error rate 88% while the truth is Down , and error rate 8% while the truth is Up

1. Now fit the logistic regression model using a training data period from 1990 to 2008, with Lag2 as the only predictor. Compute the confusion matrix and the overall fraction of correct predictions for the held out data (that is, the data from 2009 and 2010).

一張含有 文字 的圖片

自動產生的描述the result is shown on the right hand side, and overall fraction of correct increase to 62.5%.

1. Repeat (d) using LDA.

一張含有 文字 的圖片

自動產生的描述the result is shown on the right hand side, and overall fraction of correct increase to 62.5%.

1. Repeat (d) using QDA.

一張含有 文字 的圖片

自動產生的描述

This model guess “Up” in every day, so the accuracy decrease to 58.6%

1. 一張含有 文字 的圖片

   自動產生的描述Repeat (d) using KNN with K = 1.   
     
   the accuracy is 51%, it not too good.
2. 一張含有 文字 的圖片

   自動產生的描述Repeat (d) using naive Bayes.  
     
   This model guess “Up” in every day, so the accuracy decrease to 58.6%
3. Which of these methods appears to provide the best results on this data?  
     
   It seems that both LDA and generalized linear model(glm), provided highest accuracy at 62.5%.
4. Experiment with different combinations of predictors, including possible transformations and interactions, for each of the methods. Report the variables, method, and associated confusion matrix that appears to provide the best results on the held out data. Note that you should also experiment with values for K in the KNN classifier

Using the KNN method under k=3 and 5 the former is better, which have 56% accuracy, but still not as good as glm of LDA.

一張含有 文字 的圖片

自動產生的描述

Using lag 1 and 2 as predictor

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | glm | LDA | QDA | KNN k=1 | KNN k=3 | NB |
| Accuracy | 57.8 | 57.8 | 55.8 | 48 | 52 | 53.8 |

Using lag 1, lag2, volume as predictor

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | glm | LDA | QDA | KNN k=1 | KNN k=3 | NB |
| Accuracy | 52.9 | 52.9 | 46.1 | 50 | 48.1 | 42.3 |

Using lag 1 as predictor

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | glm | LDA | QDA | KNN k=1 | KNN k=3 | NB |
| Accuracy | 56.7 | 56.7 | 58.7 | 44.2 | 46.1 | 58.6 |

After several of experiment, the original model by using Lag2 as only predictor, have best accuracy. In each model, if we add or use predictor other then Lag2 the accuracy will decrease. In conclusion, the glm and LDA model using Lag2 as only predictor provide best accuracy at 62.5%.