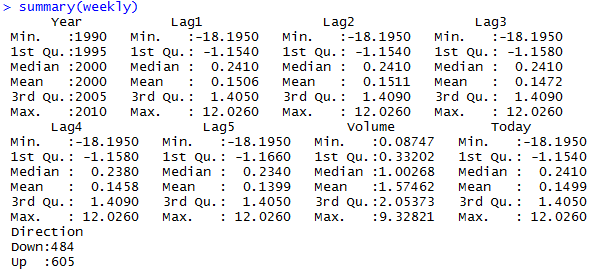
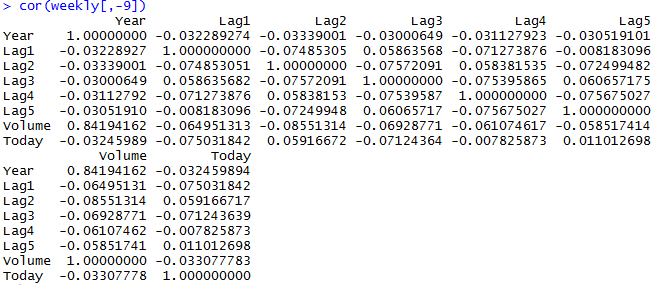
1. **This question should be answered using the Weekly data set, which is part of the ISLR package. This data is similar in nature to the Smarket data from this chapter’s lab, except that it contains 1,089 weekly returns for 21 years, from the beginning of 1990 to the end of 2010.**
2. **Produce some numerical and graphical summaries of the Weekly data. Do there appear to be any patterns?**

The below figure gives us the summary of the numerical variables of the Weekly dataset.



Let us analyze the correlation matrix for the qualitative variables in Weekly dataset

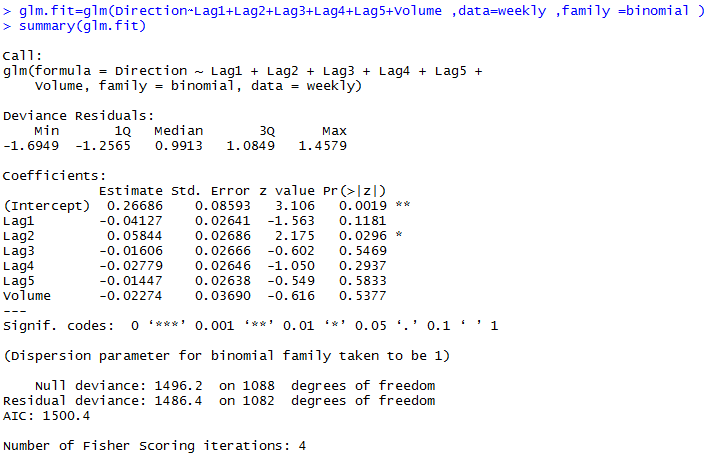


There is a high positive correlation between year and volume which indicates increase in volume of shares traded as the year’s progress. None of the Lag variables are correlated with each other.

|  |  |
| --- | --- |
|  | A closer look at the relationship between Volume and Year. We can observe that the average number of shares traded per day has increased as the year’s progressed. |

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?**

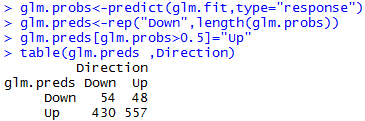
The below is the summary of the logistic regression performed on the Weekly dataset with all the Lag variables and Volume with Direction as the response.



From the summary of the Logistic Regression only Lag2 (Percentage return for 2 weeks previous) variable is statistically significant. The coefficient of Lag2 is positive, which means that when the percentage return for 2 previous week’s increases, it is more likely that Direction goes up or the market has a positive return.

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**

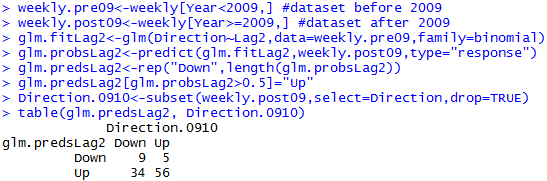
The confusion matrix is created with the below code:



The percentage of correct predictions is computed as (54+557)/1089=0.5611~56.11%. The training error rate is 43.89%. The model performs well when the market goes up 92.06% (557/(557+48)\*100) of the time. However when the market actually goes down the prediction rate falls to 11.15% (54/(430+54)\*100).

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).**

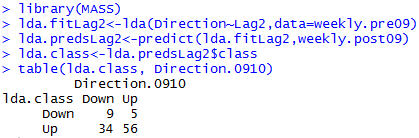
Below code is a step-by-step process to get the confusion matrix:



The fraction of correct predictions is 62.5% ((9+56/104)\*100)

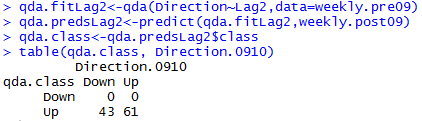
1. **Repeat (d) using LDA**

Steps below using LDA to create the confusion matrix:



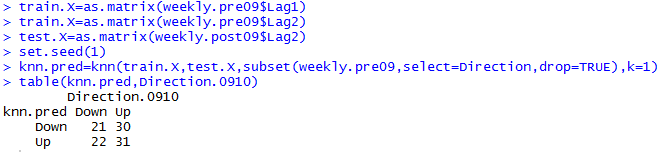
The fraction of correct predictions is 62.5% ((9+56/104)\*100)

1. **Repeat (d) using QDA**

****

The fraction of correct predictions is 58.65% ((61/104)\*100). However, the model predicts correctly only the cases where the market goes up. In the case where the market goes down the prediction error is 100% i.e. it gets the prediction wrong every time the market goes down.

1. **Repeat (d) using KNN with K = 1**



The fraction of correct predictions is 50%((52/104)\*100).

1. **Which of these methods appears to provide the best results on this data?**

By comparing the test error rates, we see that logistic regression and LDA have the same error rates, which are also the minimum error rates, followed by QDA and KNN. Hence, logistic regression and LDA provide the best results on this data

1. **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 classifiers.**

Logistic Regression:

|  |  |
| --- | --- |
|  | In logistic regression, the model showed best results when we considered Lag2 and Lag2^2. This model gave a correct prediction of 62.5 % similar to the initial model in question 1(d). The closest model to this is Direction~Lag1:Lag2 with a correct prediction rate of 58.65% |

LDA:

|  |  |
| --- | --- |
|  | In LDA, the model showed best results when we considered Lag2 and Lag2^2. This model gave a correct prediction of 61.54 % similar to the initial model in question 1(e). |

QDA:

|  |  |
| --- | --- |
|  | In QDA, the model showed best results when we considered Lag2 and Lag2^2. This model gave a correct prediction of 62.5 % similar to the initial model in question 1(f). |

KNN:

|  |  |
| --- | --- |
|  | In KNN, the model showed best results when we considered Lag2 and Lag2^2 and K = 10. This model gave a correct prediction of 59.61 % similar to the initial model in question 1(g). |

Of the four methods Logistic regression gave the best results with Lag2 and Lag2^2 as predictors with a least test error of 37.5%.

1. **Perform ROC analysis and present the results for logistic regression and LDA used for the best model chosen in Question 1(i).**

Logistic Regression:

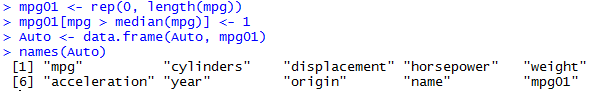
|  |  |
| --- | --- |
|  |  |

LDA:

|  |  |
| --- | --- |
|  |  |

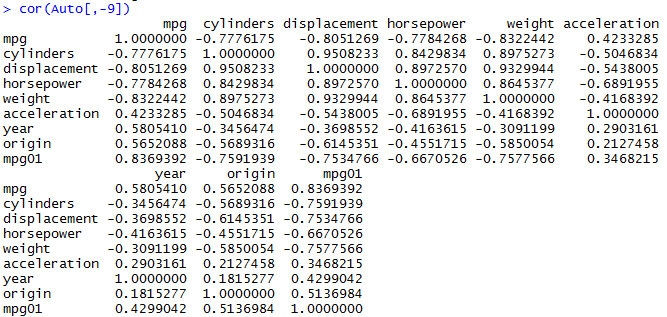
The sensitivity is for both LR and LDA 0.96, which is the percentage of times the model predicts when the market moved “Up”. However our model has a very high FPR of 0.942 and 0.948 for LR and LDA respectively. This means it wrongly predicts market going “Down” as “Up”

1. **In this problem, you will develop a model to predict whether a given car gets high or low gas mileage based on the Auto data set.**
2. **Create a binary variable, mpg01, that contains a 1 if mpg contains a value above its median, and a 0 if mpg contains a value below its median. You can compute the median using the median( ) function. Note that you may find it helpful to use the data.frame( ) function to create a single data set containing both mpg01 and the other Auto variables.**

****

1. **Explore the data graphically in order to investigate the association between mpg01 and the other features. Which of the other features seem most likely to be useful in predicting mpg01? Scatterplots and Boxplots may be useful tools to answer this question. Describe your findings.**

Finding the correlation matrix between the variables



From this correlation matrix we can see a high correlation between mpg01 and cylinders, displacement, horsepower, weight, year, acceleration

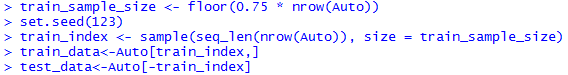
|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |

From the box plots we can derive the following conclusions:

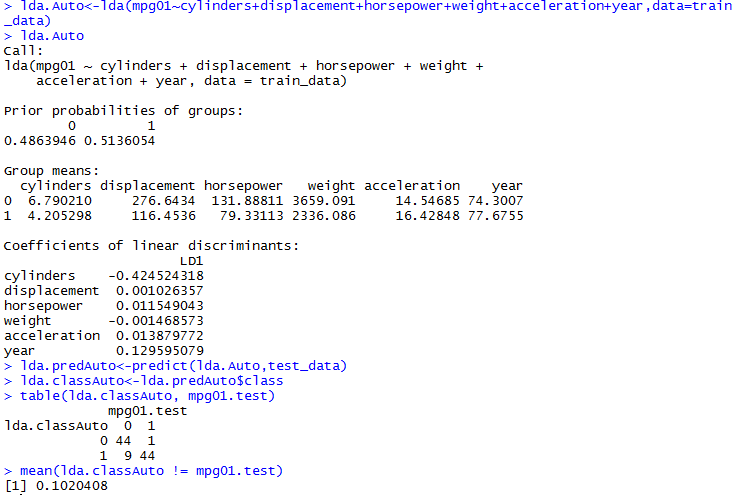
* As the number of cylinders increases mpg decreases and falls below median.
* As the engine displacement increases mpg decreases and falls below median.
* As the horsepower increases mpg decreases and falls below median.
* As the weight increases mpg decreases and falls below median.
* Higher the acceleration higher the mpg and likely to be above median.
* As the manufacturing year increases mpg increases and likely to be above median.

1. **Split the data into a training set and a test set**

Splitting the training and test data in the ratio of 75:25

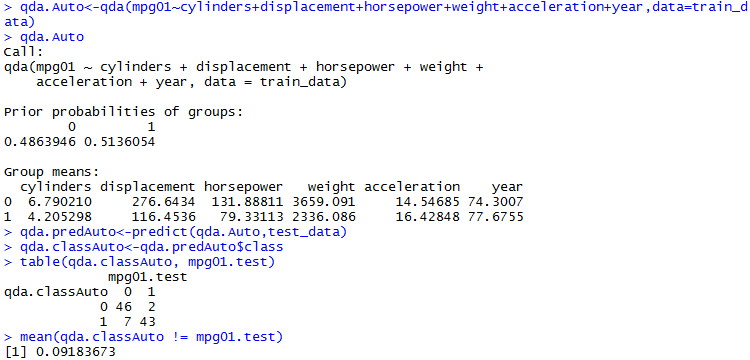


1. **Perform LDA on the training data in order to predict mpg01 using the variables that seemed most associated with mpg01 in (b). What is the test error of the model obtained?**



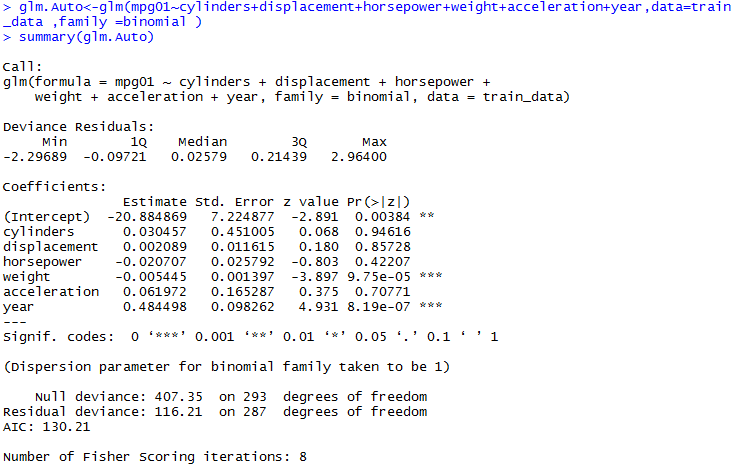
The test error of LDA is 10.20408 %

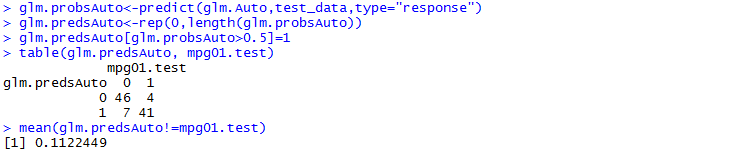
1. **Perform QDA on the training data in order to predict mpg01 using the variables that seemed most associated with mpg01 in (b). What is the test error of the model obtained?**



The test error of QDA is 9.183673%

1. **Perform logistic regression on the training data in order to predict mpg01 using the variables that seemed most associated with mpg01 in (b). What is the test error of the model obtained?**

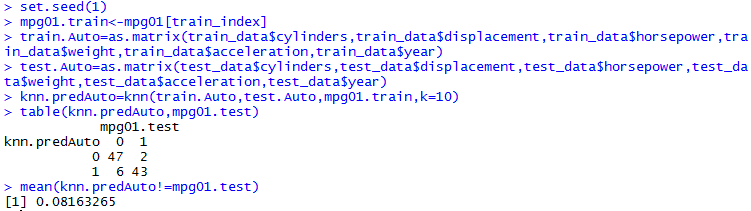
****

****

The test error of Logistic Regression is 11.22449%

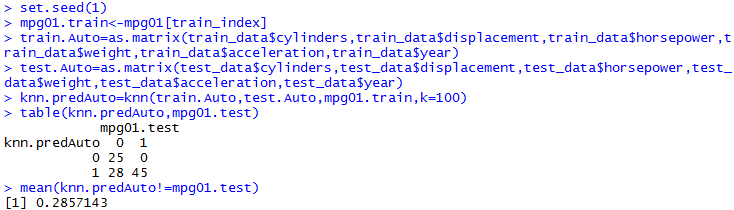
1. **Perform KNN on the training data, with several values of K, in order to predict mpg01. Use only the variables that seemed most associated with mpg01 in (b). What test errors do you obtain? Which value of K seems to perform the best on this data set?**

K=10:



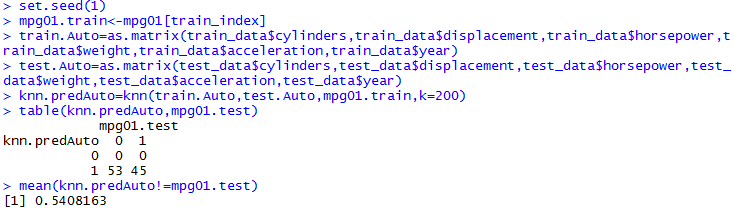
Test error: 8.163265%

K=100:



Test error: 28.57143%

K=200:



Test error: 54.08163%

K=10 seems to be the best performance. 10 nearest neighbors.