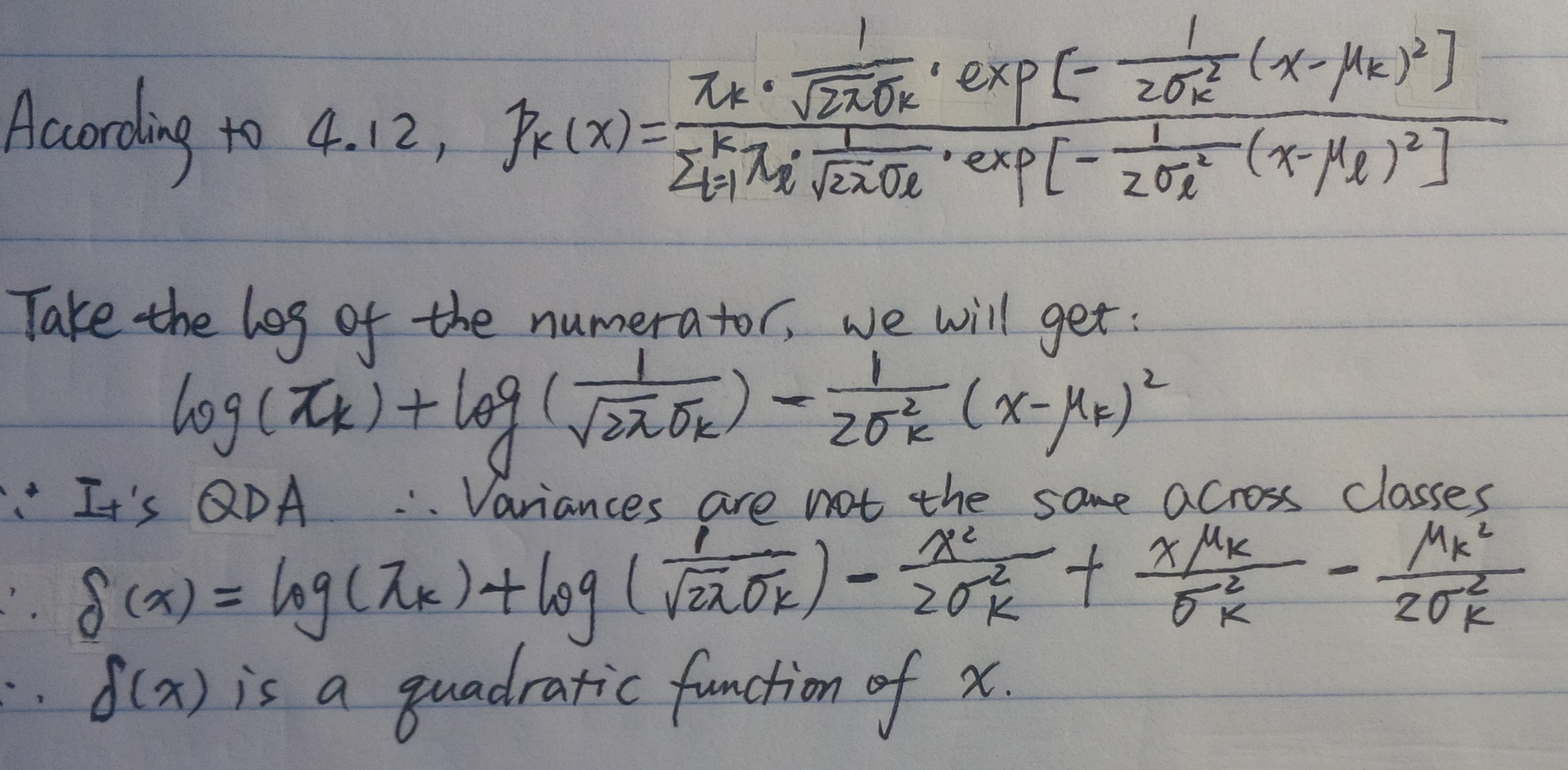
**HW 2 Chapter 4**

Luyao Zhang (NetID: lzhang94)

**Ex 3**



**Ex 4**

1. When, it uses a fraction of 10% of available observations.

When, it uses a fraction of (100x+5) % of available observations.

When x > 0.95, it uses a fraction of (105-100x) % of available observations.

Therefore, on average, to predict the response at a given X value, the fraction of available observation used is:



Therefore, on average 9.75% of available observations will be used.

1. Given that X1 and X2 are independent, we will use 9.75%\*9.75% 0.95% of the available observations to make predictions.
2. The fraction will be 
3. The drawback of KNN is that when p is very large, it’s almost that we won’t find any observations available that are near any given test observations for prediction.
4. When p=1, the length: 0.1^1=0.1; when p=2, the length: 0.1^0.50.32; when p=100, the length: 0.1^(1/100) 0.977; when we have n features (i.e., p = n, the length: 0.1^(1/n)

When the fraction of observations used is fixed, the more features we have, the longer the side of the hypercube

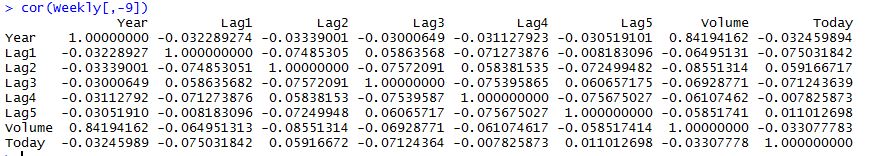
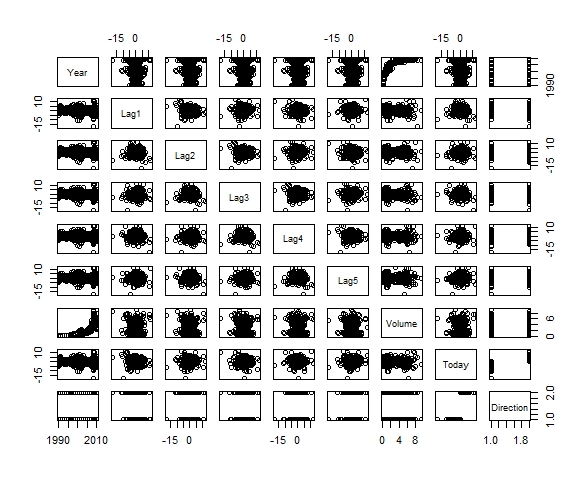
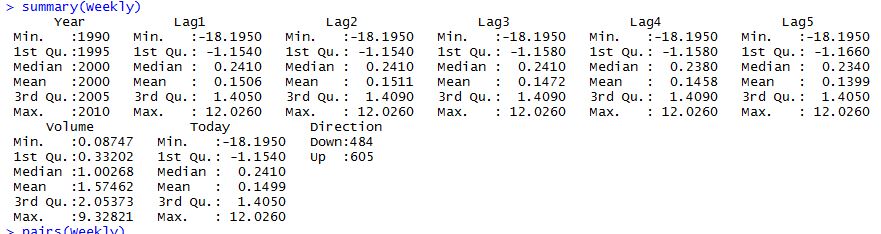
**Ex 6**

1. 
2. 

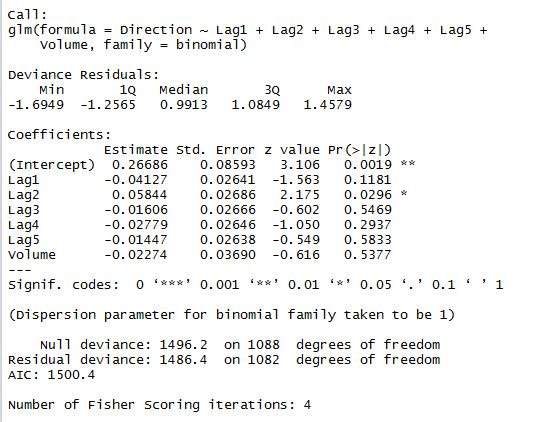


**Ex 10**

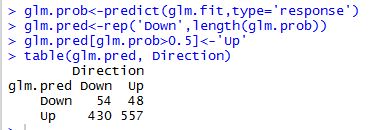
1. Some of the basic information between variables of the Weekly data is as below. According to the summary, correlations, and plots, it seems that year and volume are strongly positively correlated (r = 0.84), which means that volume increase as time goes by. Besides this, the correlation between lags and today’s returns are almost zero.



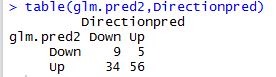
1. As shown below, Lag2 appears to be statistical significant (*p* < 0.05)



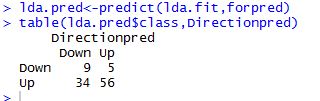
1. The confusion matrix is shown below. According to the table, the percentage of correct predictions is (54+557) / (54+48+430+557) = 56%. When the week is up, the model correctly predicts the results 557 / (48+557) = 92% of the time. When the week is down, the model is correct only 54 / (54+430) = 11% of the time. Therefore, the logistic regression model performs a lot better for up weeks than for down weeks.



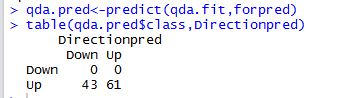
1. The confusion table is as below. According to the table, the percentage of correct prediction is (9+56) / (9+5+34+56) = 62.5%. When the week is up, the model correctly predicts the results 56 / (56+5) = 91.8% of the time. When the week is down, the model correctly predicts the results 9 / (9+34) = 20.9% of the time. Again, the model does a much better job predicting results for up weeks than for down weeks.



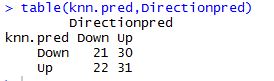
1. The LDA results are as below. Turns out the percentage of the model correctly predicting the results is 62.5%. For weeks that are up, the model predicts correctly 91.8% of the time, while for weeks that are down, the correct rate is only 20.9%. This is very close to what we obtained using logistic regression.



1. The QDA results are as below. According to the table, the percentage of the model correctly predicting the results is about 58.7%. For weeks that are up, the model is correct 100% of the time, but when the market is down, the model is never correct! Even though the model basically classifies weeks to “Up” all the time, the model still has achieved an overall correct rate as high as 58.7%.

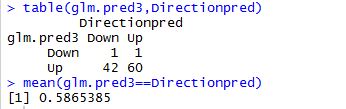


1. The results using KNN is as below. According to the table, the percentage of the model correctly predicting the results is 50%. For up weeks, the model is correct 50.8% of the time, while for down weeks, the model is correct 48.8% of the time.

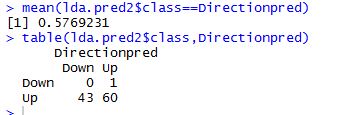


1. In terms of overall correct rate (or error rate), LDA and logistic regression have similarly low error rate. QDA and KNN have higher error rates, and therefore, are not as good as LAD and logistic regression.
2. Results of models using different methods with different variable combinations are as below. Compared with results obtained from questions above, the original logistic regression and LDA have the lowest error rate (37.5%).

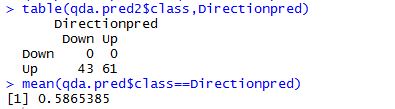
Logistic regression with the interaction of Lag1 and Lag 2



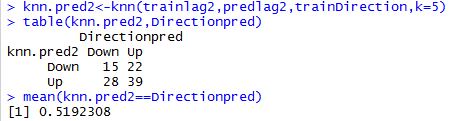
LDA with the interaction of Lag1 and Lag2



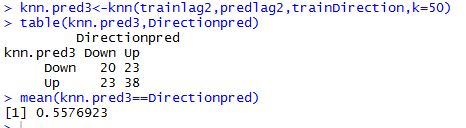
QDA with Lag2 and (Lag2) ^2



KNN with k=5

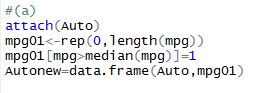


KNN with k=50



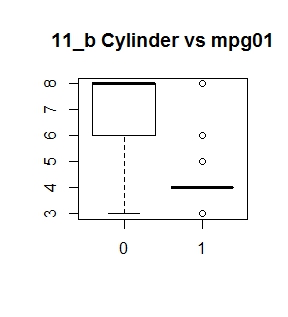
**EX 11**

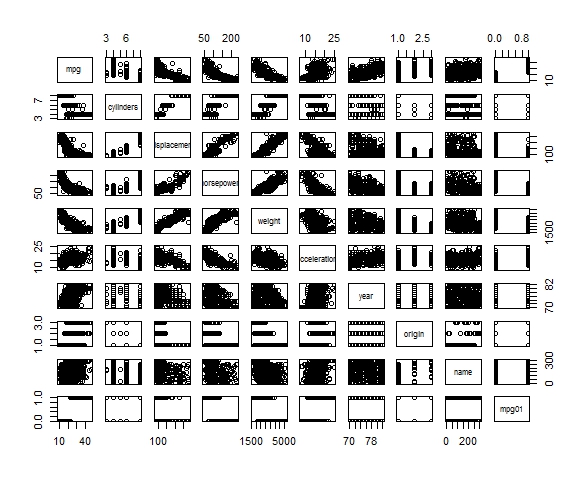
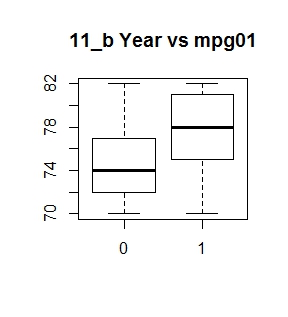
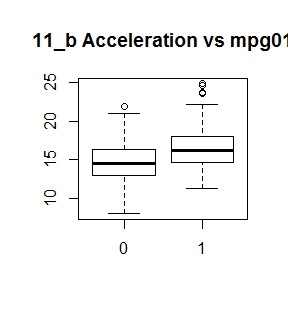
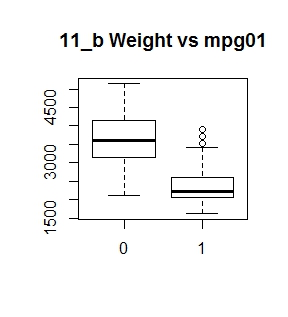
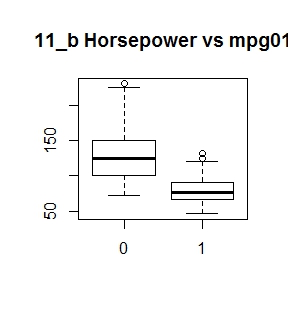
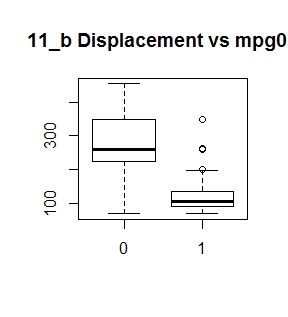
1. The code for creating the binary data and the summary of the new binary variable are as below:



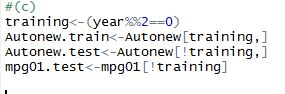
C:\Users\lzhang94\AppData\Local\Microsoft\Windows\INetCache\Content.Word\11_a_2.jpg

1. Plots are as below. The scatterplots don’t contain much information. It looks like cylinders, horsepower, displacement, and weight can be useful predictors of mpg01.

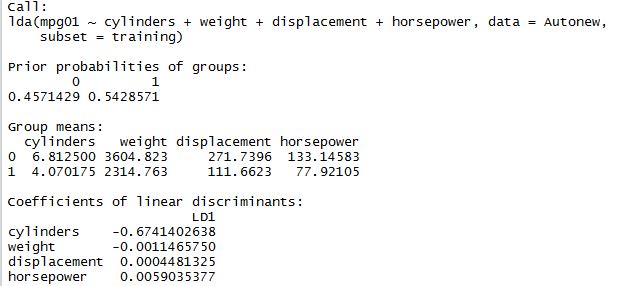


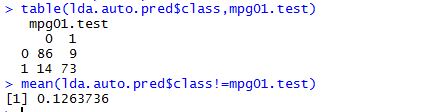


1. I split the data based on year (odd and even). The code is as below:

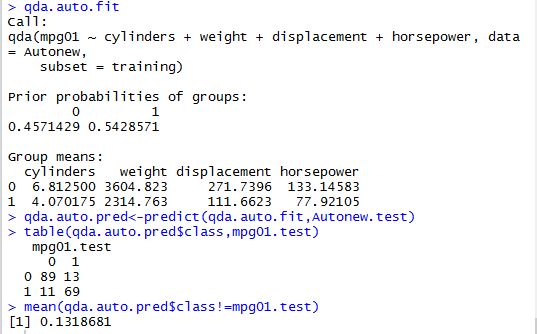


1. The summary of the model is as below. The error rate is 12.64%.

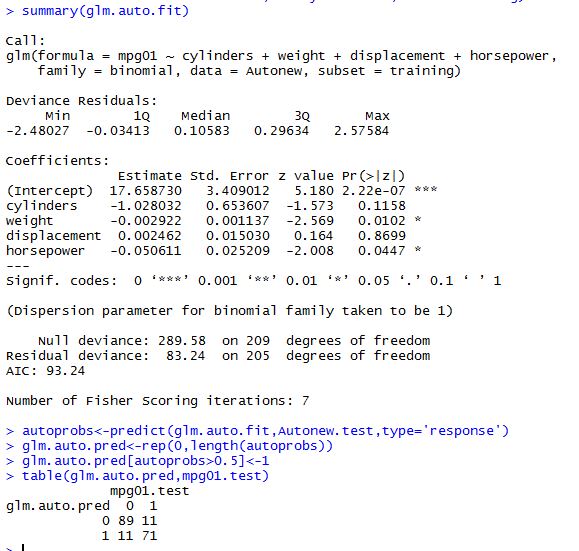




1. The summary of the QDA model is as below. The error rate is 13.19%.

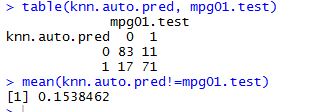


1. The summary of logistic regression and confusion matrix are as below. The error rate is 12.09%.

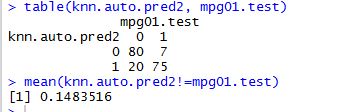


1. K=1, 50, and 100 were examined respectively, and the results can be found below. Looks like when K=100, the model has a lower error rate (14.29%), compared with when K=1 (error rate = 15.38%) or when K = 50 (error rate = 14.84%).

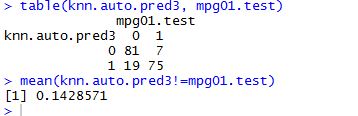
**K=1**



**K=50**



**K=100**

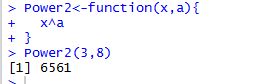


**Ex 12**

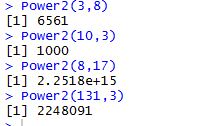
1. The result of the function Power () is as below:



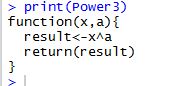
1. Power2 () is as below:



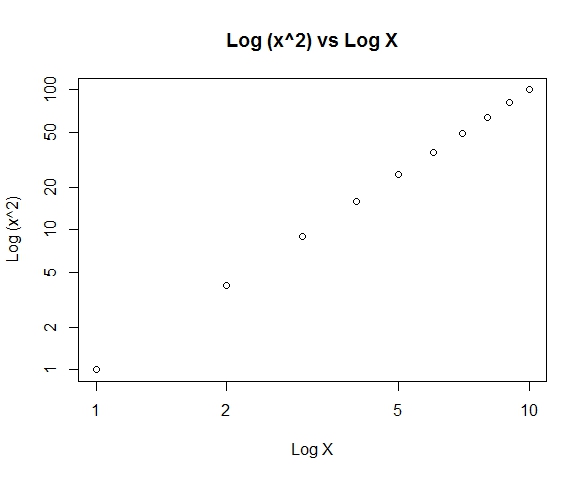
1. The results are as below:



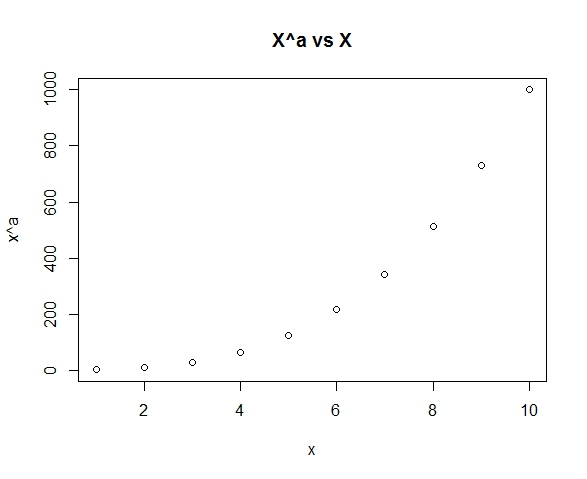
1. The function Power3 () should look like:



1. The plot is as below:



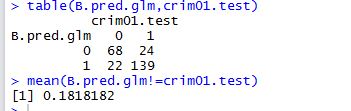
1. The plot obtained via PlotPower when x=c(1:10) is as below:



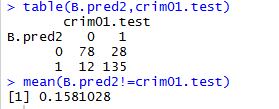
**Ex 13**

**The first 253 observations serve as the training data, while the rest 253 observations serve as the test data.**

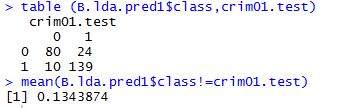
1. Fit a logistic regression model using all predictors except crim01 and crim. The overall error rate is 18.18%



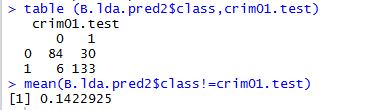
1. Fit a logistic regression model using all predictors but crim01, crim, nox, and tax. The overall error rate is 15.81%.



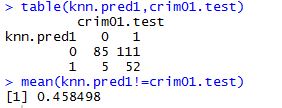
1. LDA using all predictors but crim and crim01. The overall error rate is 13.44%



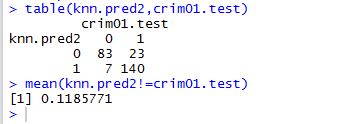
1. LDA using all predictors but crim, crim01, nox, and tax. The overall error rate is 14.23%.



1. KNN with k = 1. Overall error rate is 45.85%.



1. KNN with k = 10. Overall error rate is 11.86%.



1. KNN with k = 100. Overall error rate is 49.01%.

