Homework 3

Andrew Boschee

2/7/2020

No collaborators Outside Resources: rdocumentation.com

Question 4.7.1, pg 168 - Using a little bit of algebra, prove that (4.2) is equivalent to (4.3). In other words, the logistic function representation and the logit representation for the logistic regression model are equivalent.

Logistic Function:

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}$$

$$\frac{1}{p(X)} = \frac{1+e^{\beta_0+\beta_1 X}}{e^{\beta_0+\beta_1 X}}$$

$$\frac{1}{p(X)} = \frac{1}{e^{\beta_0+\beta_1 X}} + \frac{e^{\beta_0+\beta_1 X}}{e^{\beta_0+\beta_1 X}}$$

$$\frac{1}{p(X)} = 1 + \frac{1}{e^{\beta_0+\beta_1 X}}$$

$$e^{\beta_0+\beta_1 X} = \frac{p(X)}{1-p(X)}$$

Question 4.7.10, pg 171 - This question should be answered using teh 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, exceot that it contains 1,089 weekly returns for 21 years, from the beginning of 1990 to the end of 2010.

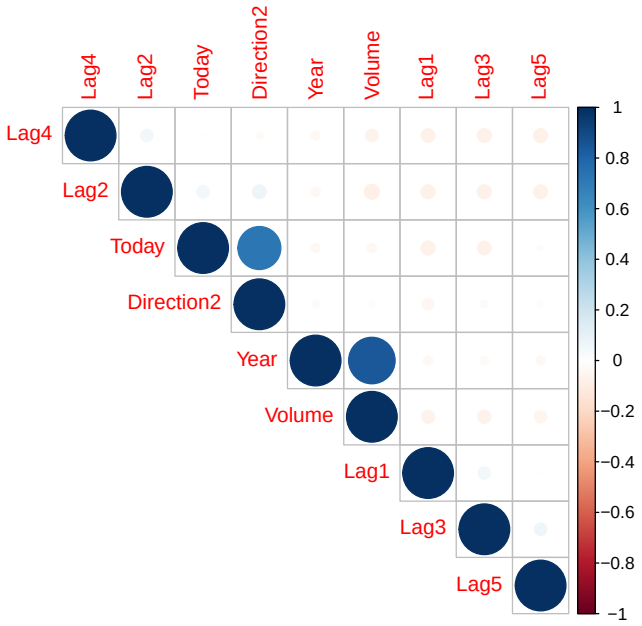
a. Produce some numerical and graphical summaries of the Weekly data. Do there appear to be any patterns?

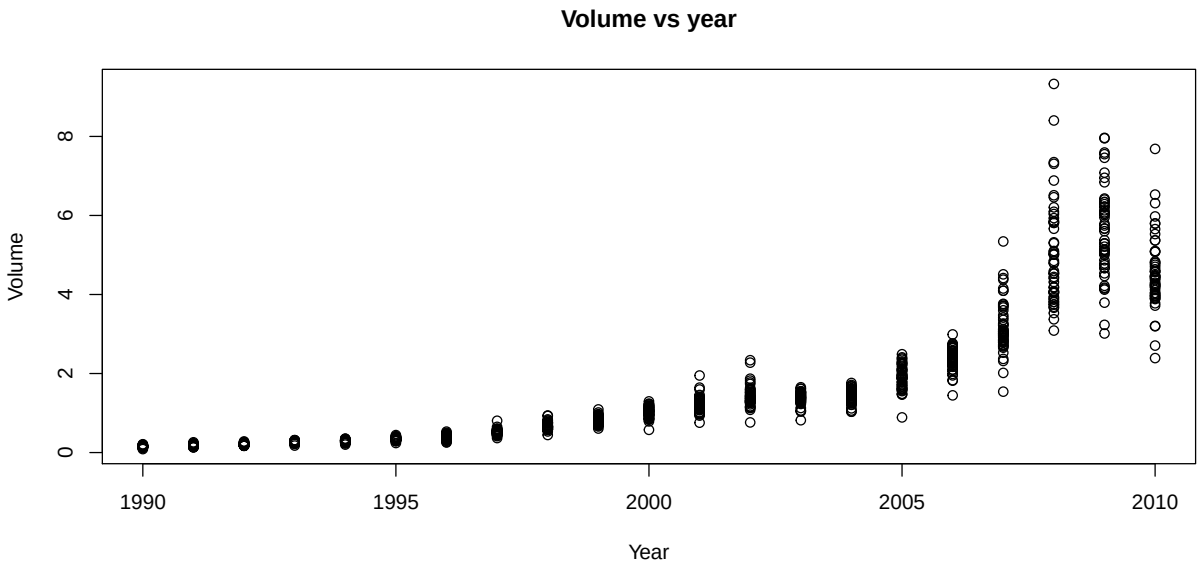
```
##       Year           Lag1               Lag2               Lag3         
##  Min.   :1990   Min.   :-18.1950   Min.   :-18.1950   Min.   :-18.1950  
##  1st Qu.:1995   1st Qu.: -1.1540   1st Qu.: -1.1540   1st Qu.: -1.1580  
##  Median :2000   Median :  0.2410   Median :  0.2410   Median :  0.2410  
##  Mean   :2000   Mean   :  0.1506   Mean   :  0.1511   Mean   :  0.1472  
##  3rd Qu.:2005   3rd Qu.:  1.4050   3rd Qu.:  1.4090   3rd Qu.:  1.4090  
##  Max.   :2010   Max.   : 12.0260   Max.   : 12.0260   Max.   : 12.0260  
##       Lag4               Lag5              Volume       
##  Min.   :-18.1950   Min.   :-18.1950   Min.   :0.08747  
##  1st Qu.: -1.1580   1st Qu.: -1.1660   1st Qu.:0.33202  
##  Median :  0.2380   Median :  0.2340   Median :1.00268  
##  Mean   :  0.1458   Mean   :  0.1399   Mean   :1.57462  
##  3rd Qu.:  1.4090   3rd Qu.:  1.4050   3rd Qu.:2.05373  
##  Max.   : 12.0260   Max.   : 12.0260   Max.   :9.32821  
##      Today          Direction 
##  Min.   :-18.1950   Down:484  
##  1st Qu.: -1.1540   Up  :605  
##  Median :  0.2410             
##  Mean   :  0.1499             
##  3rd Qu.:  1.4050             
##  Max.   : 12.0260
```

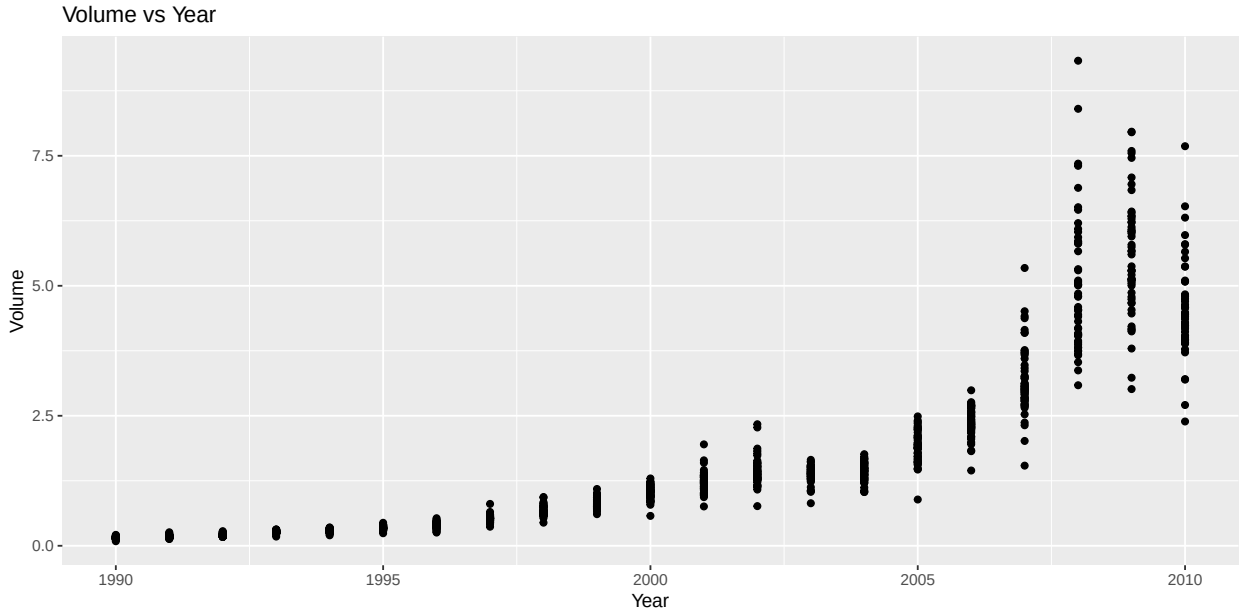
Table 1: Correlation Table

	Year	Lag1	Lag2	Lag3	Lag4	Lag5	Volume	Today	Direction2
Year	1.0000	-0.0323	-0.0334	-0.0300	-0.0311	-0.0305	0.8419	-0.0325	-0.0222
Lag1	-0.0323	1.0000	-0.0749	0.0586	-0.0713	-0.0082	-0.0650	-0.0750	-0.0500
Lag2	-0.0334	-0.0749	1.0000	-0.0757	0.0584	-0.0725	-0.0855	0.0592	0.0727
Lag3	-0.0300	0.0586	-0.0757	1.0000	-0.0754	0.0607	-0.0693	-0.0712	-0.0229
Lag4	-0.0311	-0.0713	0.0584	-0.0754	1.0000	-0.0757	-0.0611	-0.0078	-0.0205
Lag5	-0.0305	-0.0082	-0.0725	0.0607	-0.0757	1.0000	-0.0585	0.0110	-0.0182
Volume	0.8419	-0.0650	-0.0855	-0.0693	-0.0611	-0.0585	1.0000	-0.0331	-0.0180
Today	-0.0325	-0.0750	0.0592	-0.0712	-0.0078	0.0110	-0.0331	1.0000	0.7200
Direction2	-0.0222	-0.0500	0.0727	-0.0229	-0.0205	-0.0182	-0.0180	0.7200	1.0000

From the tables given, Lag2 and Lag4 seem to show the most correlation among the variables. Can see that there is negative correlation among all of the other lag variables.







We can see that the volume has drastically increase in recent years as technology and high frequency trading has evolved. Interesting to see how this changes in upcoming years as brokerage firms are no longer charging transaction fees for trades. Another thing that comes to mind besides technological advances is the economy in general as the frequency increased up until 2010 eventually going down as the economy struggled.

b. Use the fully 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 only variable that appears to be significant is Lag2. This comes as a bit of a surprise as I was thinking that Volume would most likely stand out in this scenario.

##

```
## Call:
## glm(formula = Direction ~ Lag1 + Lag2 + Lag3 + Lag4 + Lag5 +
##     Volume, family = binomial, data = Weekly)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6949  -1.2565   0.9913   1.0849   1.4579
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.26686    0.08593   3.106   0.0019 **
## Lag1        -0.04127    0.02641  -1.563   0.1181
## Lag2         0.05844    0.02686   2.175   0.0296 *
## Lag3        -0.01606    0.02666  -0.602   0.5469
## Lag4        -0.02779    0.02646  -1.050   0.2937
## Lag5        -0.01447    0.02638  -0.549   0.5833
## Volume      -0.02274    0.03690  -0.616   0.5377
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1496.2  on 1088  degrees of freedom
## Residual deviance: 1486.4  on 1082  degrees of freedom
## AIC: 1500.4
##
## Number of Fisher Scoring iterations: 4
```

Table 2: P-Values of Predictors

	P-Value
(Intercept)	0.0018988
Lag1	0.1181444
Lag2	0.0296014
Lag3	0.5469239
Lag4	0.2936533
Lag5	0.5833482
Volume	0.5376748

c. 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.

We can see that the model was very optimistic guessing 987 weeks up and 102 down in comparison to the 604/484 that actually occurred. The confusion matrix shows that there were many false positives as a result of this in the top-right column where there were 430 misclassifications of going up but instead went down. There were 48 false negatives, 557 true positives, and 54 true negatives. This resulted in an unimpressive 56 % accuracy

Table 3: Prediction Confusion Matrix

	Down	Up
Down	54	430
Up	48	557

Table 4: Predictions

	x
Down	102
Up	987

Table 5: Actual

	x
Down	484
Up	605

Table 6: Accuracy of Predictions

Accuracy
0.5610652

d. 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).

Table 7: Test Set Confusion Matrix

	Down	Up
Down	9	34
Up	5	56

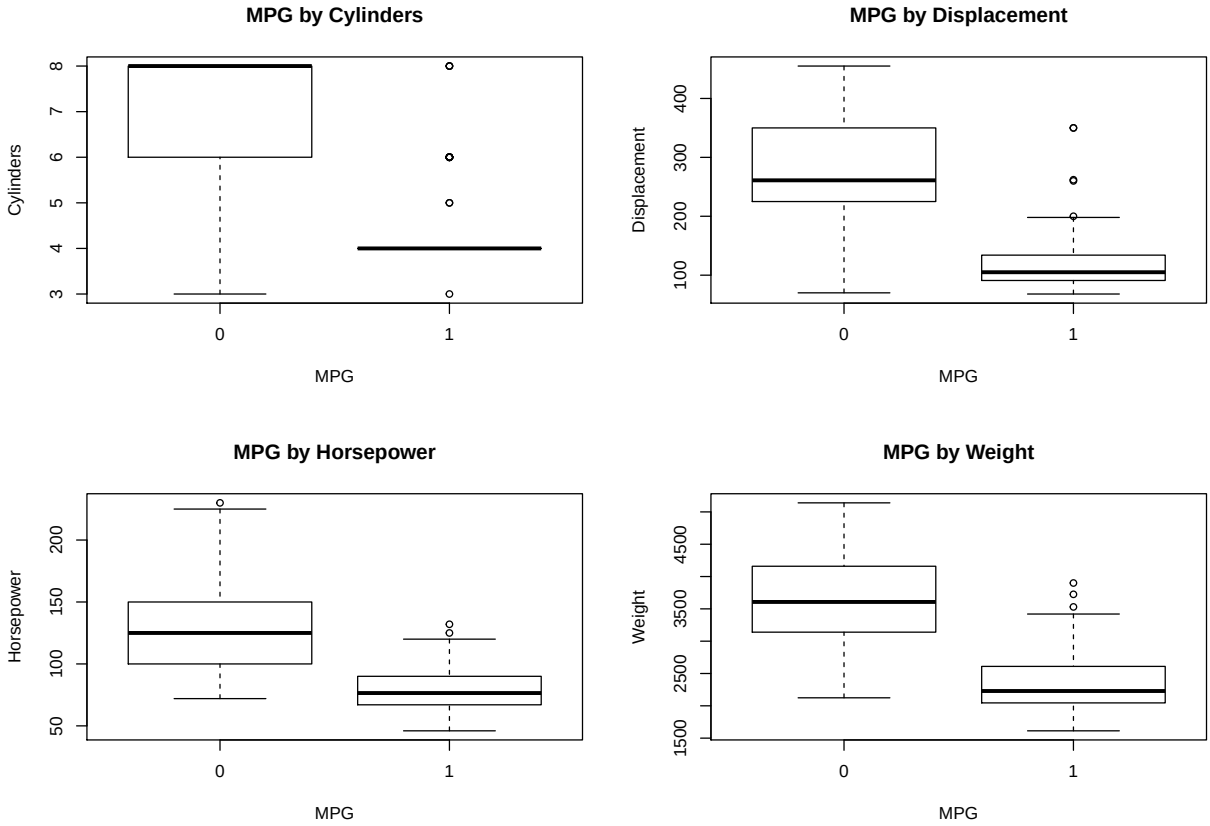
Question 4.7.11, pg 172 - In this problem, you will develop a model to predict whether a given car gets high or low gas mileage based on the Auto data set.

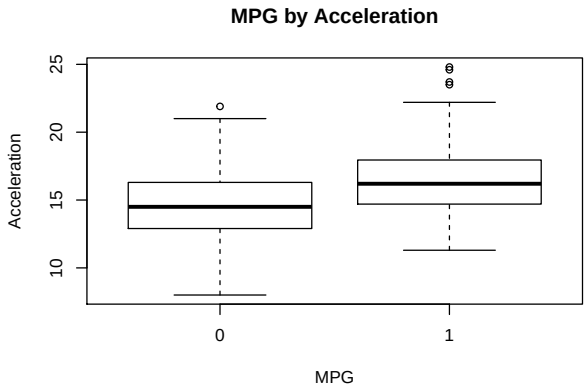
a. Create a binary variable mpg01, that contains a 1 if mpg contains a value above its median, and a 0 if mpg contains a value below the median.

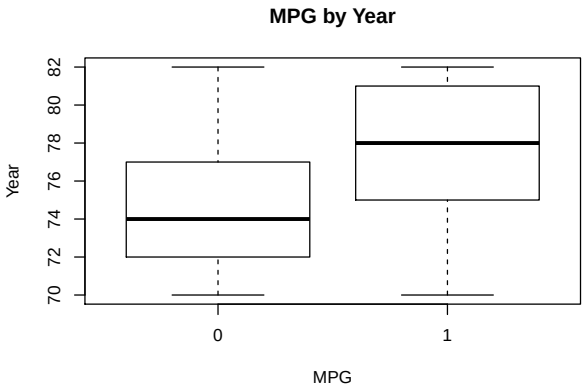
Created variable as factor using ifelse function comparing median mpg to mpg of each row.

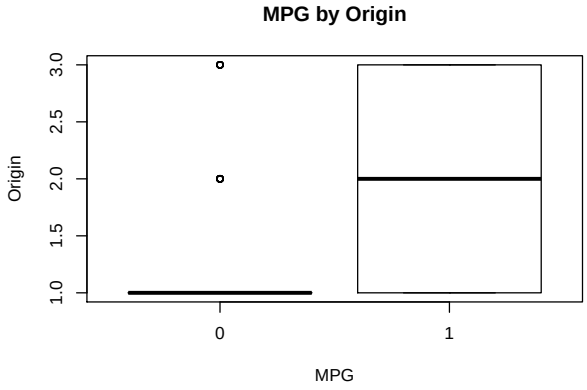
b. Explore the data set graphically in order to investigate the association between mpg01 and the other features. Which of the other features seem most likely to be useful in predicting mgp01? Scatterplots and boxplots may be useful tools to answer this question. Describe your findings.

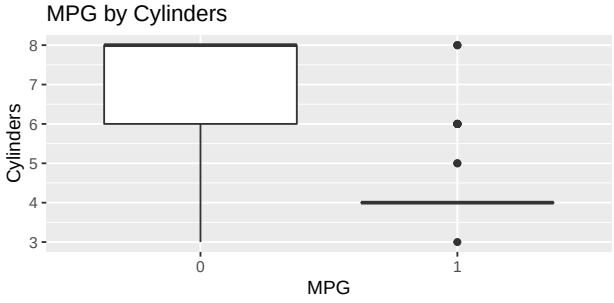
There were a few predictors that surprised me regarding there mpg being above the median. Acceleration is the first thing that caught my eye and there are a couple other factors that we don't see that I would be interested. Believe there are some correlations between many of these variables that can contribute to this difference (year as newer cars are more efficient, weight, hp). Second, horsepower had very wide IQR for not above median and very narrow IQR for those above the median as well as cylinders and displacement.

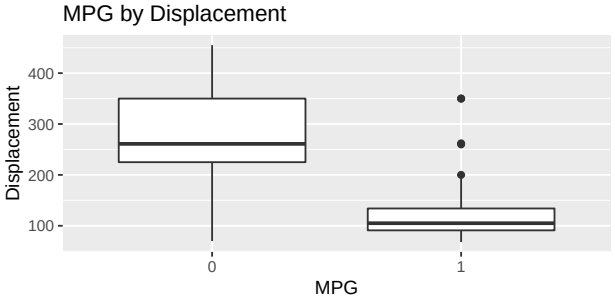


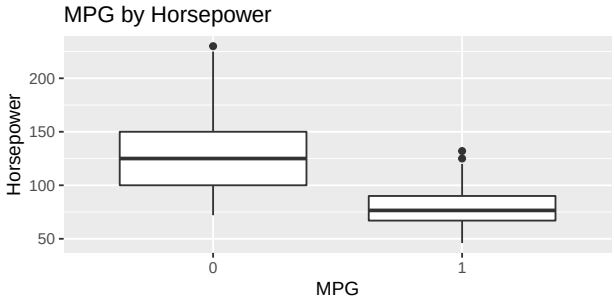

MPG by Acceleration
Acceleration
10
15
20
25
0
1
MPG

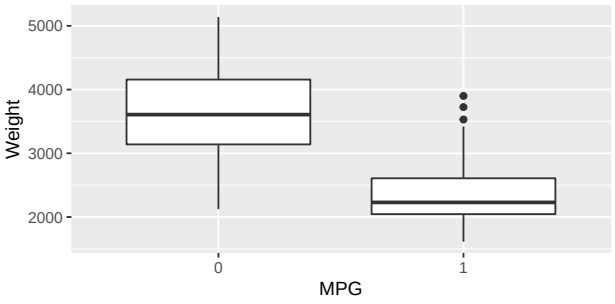

MPG by Year
Year
70
72
74
76
78
80
82
0
1
MPG

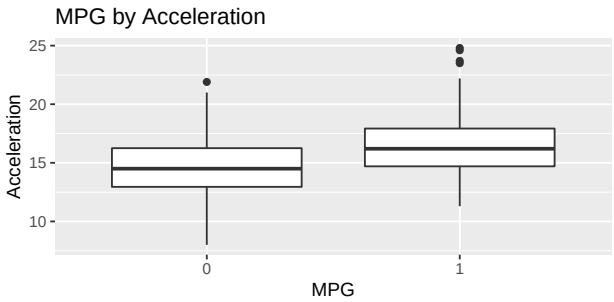

MPG by Origin
Origin
1.0
1.5
2.0
2.5
3.0
0
1
MPG

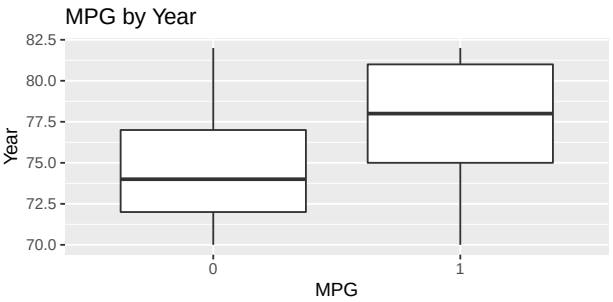

MPG by Cylinders
Cylinders
3
4
5
6
7
8
0
1
MPG

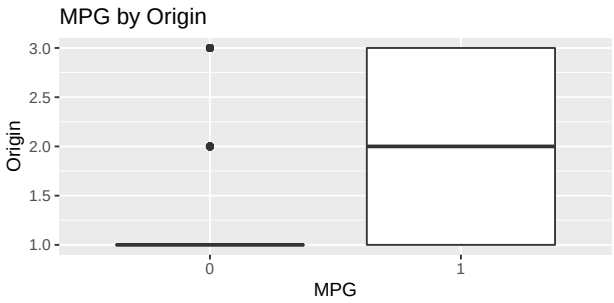

MPG by Displacement
Displacement
100
200
300
400
0
1
MPG


MPG by Horsepower
Horsepower
50
100
150
200
0
1
MPG


Weight
2000
3000
4000
5000
0
1
MPG







c. Split the data into training and test sets.

Using a 75/25 split, the count of rows for the training and test sets are shown below.

Table 8: Row Count of Each DataSet (75/25)

Auto Total Row Count	Train Set Row Count	Test Set Row Count
392	294	98

f. 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?

See first column of table 10 at the bottom

Question 4 - Write a function in RMD that calculates the misclassification rate, sensitivity, and specificity. The inputs for this function are a cutoff point, predicted probabilities, and original binary response. Test your function using the model from 4.7.10 b. (This needs to be an actual function using the function() command, not just a chunk of code). This will be something you will want to use throughout the semester, since we will be calculating these a lot! *Show the function code you wrote in your final write-up.*

Creating the function with three parameters (threshold, predictedProb, binaryResp), we can easily give this input to receive the misclassification rate, sensitivity, and specificity. First the function creates an empty list for the results. Second, the prediction output is compared to desired cutoff and assigned binary value and stored in predictions variable. Third, confusion matrix is created from binary response parameter and predictions created. Calculations are then made for each output and stored in results before being turned into a dataframe and returned as output.

```r
# accept three inputs to calculate misclassification, sensitivity, and specificity
classificationSummary <- function(threshold, predictedProb, binaryResp) {
    output <- list()
    predictions <- ifelse(predictedProb > threshold, 1, 0)
    confMatrix <- table(binaryResp, predictions)
    output$misclassificationRate <- 1- ((confMatrix[1,1] +confMatrix[2,2])/
                                        (confMatrix[1,1] +
                                         confMatrix[1,2]+confMatrix[2,1] +
                                         confMatrix[2,2]))

    output$sensitivity <- confMatrix[2,2]/(confMatrix[2,2] + confMatrix[2,1])
    output$specificity <- confMatrix[1,1]/(confMatrix[1,1] + confMatrix[1,2])
    return(as.data.frame(output))
}

# store values from function for question
classifSumm <- classificationSummary(0.5, weeklyResp, Weekly$Direction)
```

The classification summary shows the unimpressive classification rate along with the high sensitivity and low specificity. Adjusting the threshold may give better results in this situation, but would not be highly optimistic when it comes to predicting market returns.

Table 9: Classification Summary

Misclassification Rate	Sensitivity	Specificity
0.4389348	0.9206612	0.1115702

Question 3, Part f from above

Using the function created in question 4, I applied it to the MPG with the arguments (0.5, autoPred, test$mpg01). Thsi gives us the misclassification rate, sensitivity and specificity. The misclassification rate was very impressive in comparison to the market predictions only misclassifying about 10%.

This doesn't come as much of a surprise since the market has so many undefined factors and cars are a very standardized object.

Table 10: MPG Classification Summary

Misclassification Rate	Sensitivity	Specificity
0.122449	0.893617	0.8627451