Modern Applied Statistics exercises from ISLR

Yamuna Dhungana

Libraries required for the assignment

Exercises (ISLR)

1. 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 logit representation for the logistic regression model are equivalent

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

$$P(X) + P(X)(e^{\beta_0+\beta_1 X}) - e^{\beta_0+\beta_1 X} = 0$$

$$P(X)(e^{\beta_0+\beta_1 X}) - e^{\beta_0+\beta_1 X} = -P(X)$$

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

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

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

Hence proved.

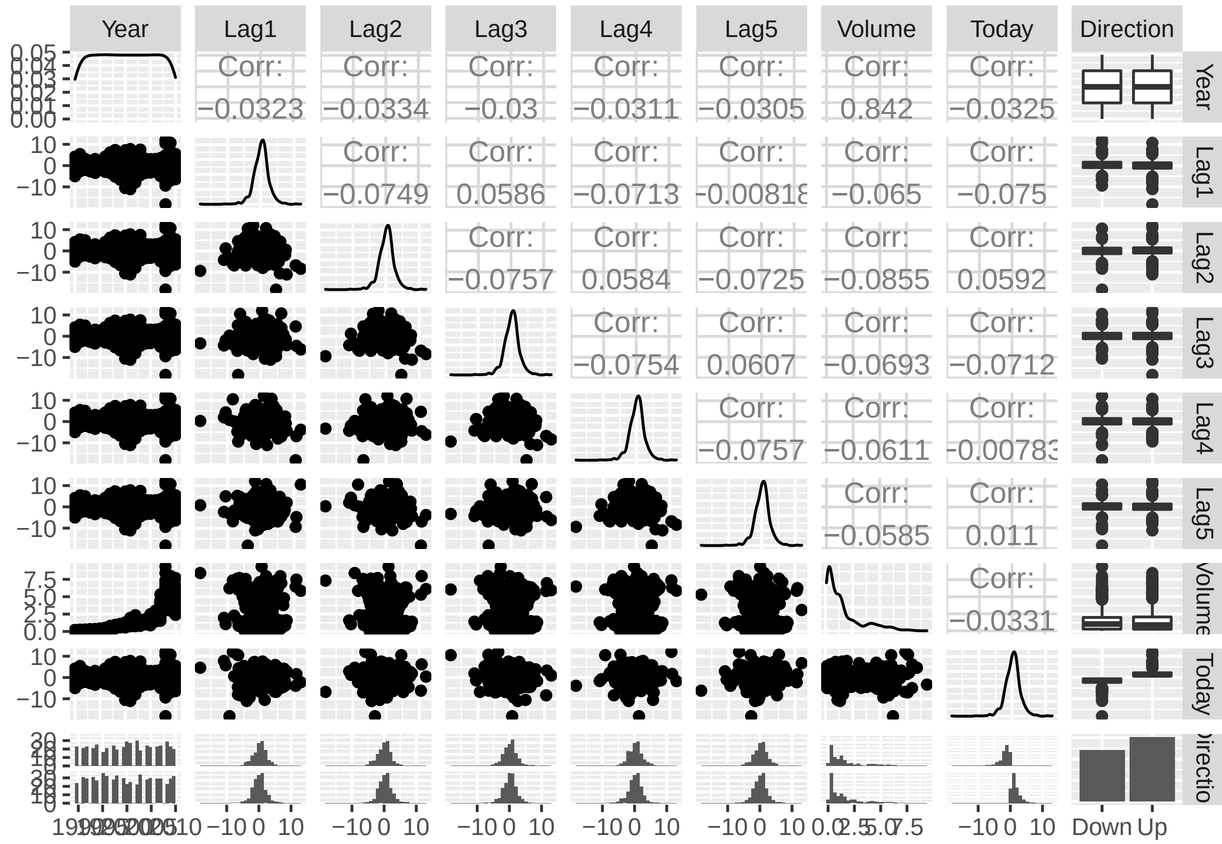
2. Question 4.7.10(a-d) pg 171 This question should be answered using the Weekly data set, which is part of the ISLR package. T his data is similar in nature to the Smarket data from this chapter's lab, except that it contains 1, 089 weekly returns for 21 years, from the beginning of 1990 to the end of 2010.

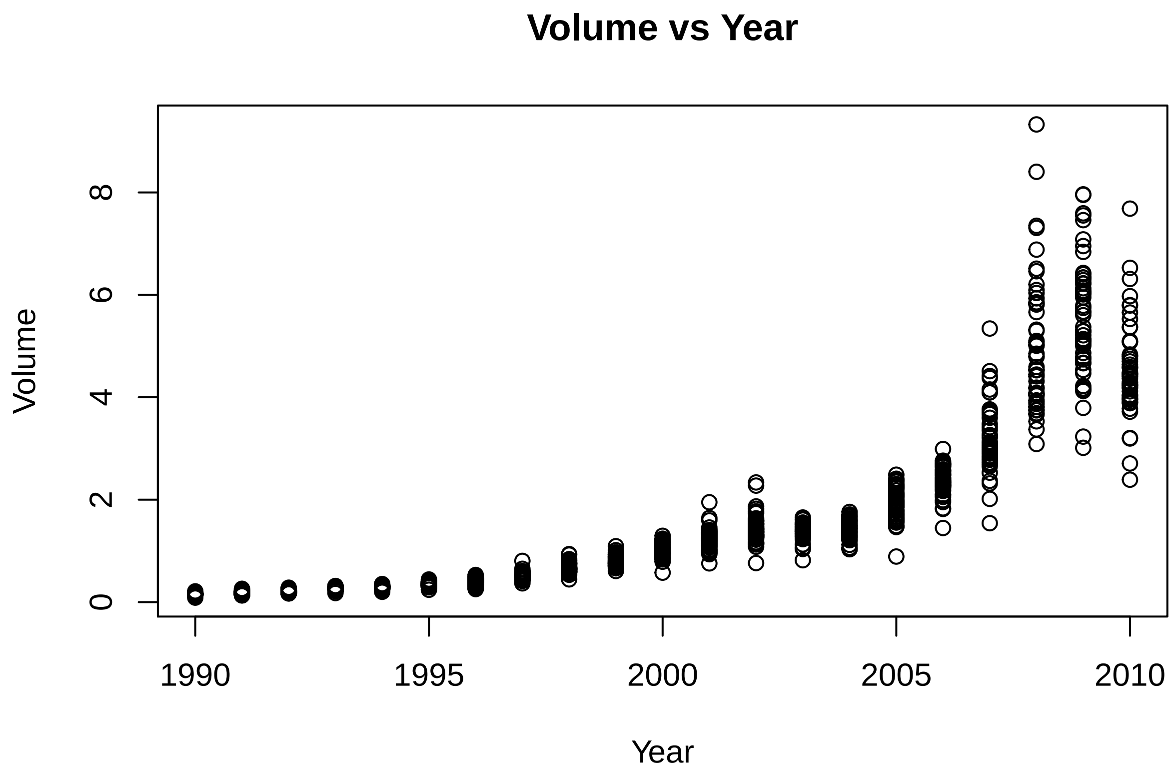
a. Produce some numerical and graphical summaries of the Weekly data. Do there appear to be any patterns?

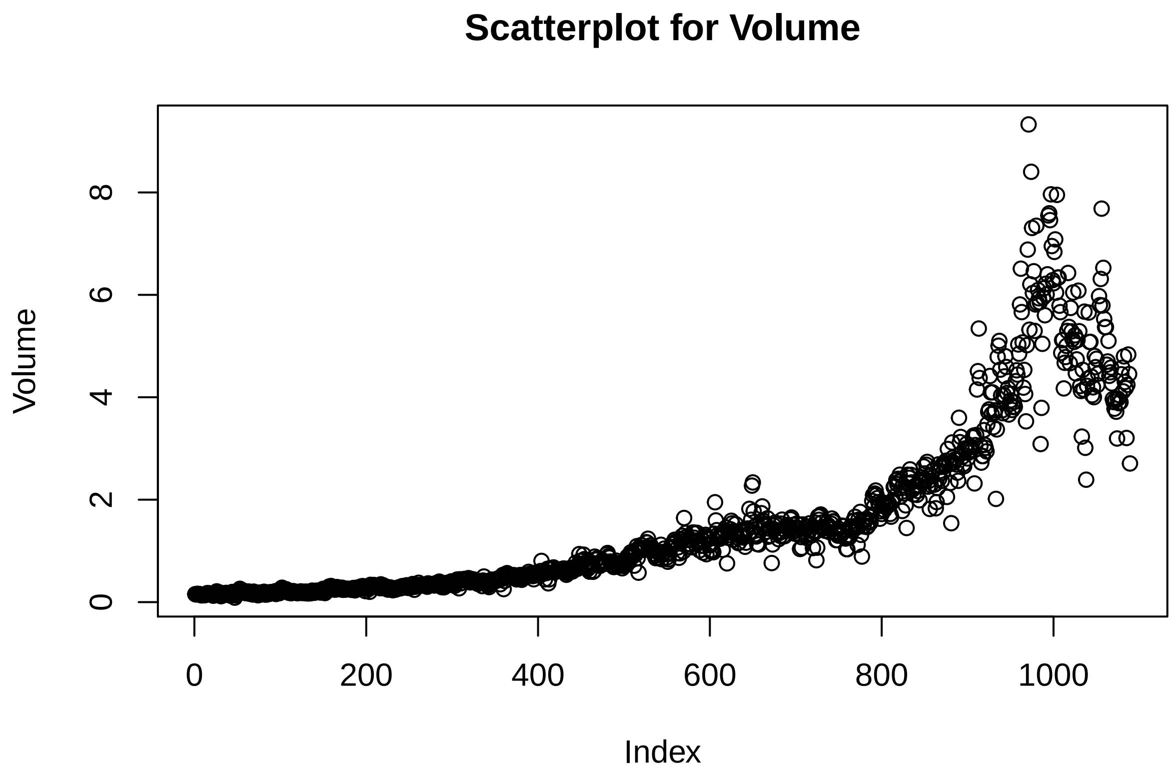
```
##       Year          Lag1               Lag2               Lag3         
##  Min.   :1990   Min.   :-18.1950   Min.   :-18.1950   Min.   :-18.1950  
##  1st Qu.:1995   1st Qu.: -1.1540   1st Qu.: -1.1540   1st Qu.: -1.1580  
##  Median :2000   Median :  0.2410   Median :  0.2410   Median :  0.2410  
##  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            Today         
##  Min.   :-18.1950   Min.   :-18.1950   Min.   :0.08747   Min.   :-18.1950  
```

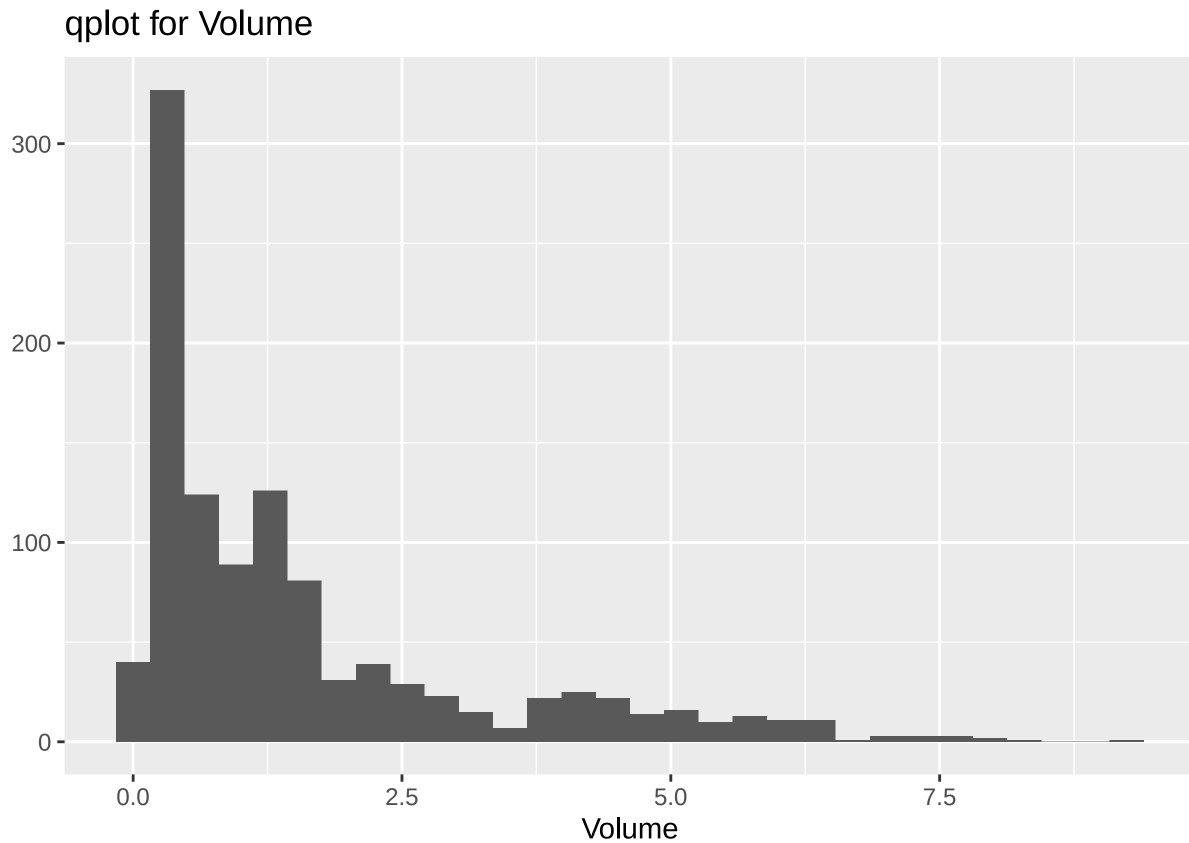
```
##  1st Qu.: -1.1580   1st Qu.: -1.1660   1st Qu.:0.33202   1st Qu.: -1.1540
##  Median :  0.2380   Median :  0.2340   Median :1.00268   Median :  0.2410
##  Mean   :  0.1458   Mean   :  0.1399   Mean   :1.57462   Mean   :  0.1499
##  3rd Qu.:  1.4090   3rd Qu.:  1.4050   3rd Qu.:2.05373   3rd Qu.:  1.4050
##  Max.   : 12.0260   Max.   : 12.0260   Max.   :9.32821   Max.   : 12.0260
##  Direction
##  Down:484
##  Up  :605
##
##
##
##
```

```
##               Year         Lag1        Lag2        Lag3         Lag4
## Year    1.00000000 -0.032289274 -0.03339001 -0.03000649 -0.031127923
## Lag1   -0.03228927  1.000000000 -0.07485305  0.05863568 -0.071273876
## Lag2   -0.03339001 -0.074853051  1.00000000 -0.07572091  0.058381535
## Lag3   -0.03000649  0.058635682 -0.07572091  1.00000000 -0.075395865
## Lag4   -0.03112792 -0.071273876  0.05838153 -0.07539587  1.000000000
## Lag5   -0.03051910 -0.008183096 -0.07249948  0.06065717 -0.075675027
## Volume  0.84194162 -0.064951313 -0.08551314 -0.06928771 -0.061074617
## Today  -0.03245989 -0.075031842  0.05916672 -0.07124364 -0.007825873
##                Lag5      Volume        Today
## Year   -0.030519101  0.84194162 -0.032459894
## Lag1   -0.008183096 -0.06495131 -0.075031842
## Lag2   -0.072499482 -0.08551314  0.059166717
## Lag3    0.060657175 -0.06928771 -0.071243639
## Lag4   -0.075675027 -0.06107462 -0.007825873
## Lag5    1.000000000 -0.05851741  0.011012698
## Volume -0.058517414  1.00000000 -0.033077783
## Today   0.011012698 -0.03307778  1.000000000
```


Year
Lag1
Lag2
Lag3
Lag4
Lag5
Volume
Today
Direction
Corr:
-0.0323
Corr:
-0.0334
Corr:
-0.03
Corr:
-0.0311
Corr:
-0.0305
Corr:
0.842
Corr:
-0.0325
Corr:
-0.0749
Corr:
0.0586
Corr:
-0.0713
Corr:
-0.00818
Corr:
-0.065
Corr:
-0.075
Corr:
-0.0757
Corr:
0.0584
Corr:
-0.0725
Corr:
-0.0855
Corr:
0.0592
Corr:
-0.0754
Corr:
0.0607
Corr:
-0.0693
Corr:
-0.0712
Corr:
-0.0757
Corr:
-0.0611
Corr:
-0.00783
Corr:
-0.0585
Corr:
0.011
Corr:
-0.0331
Down
Up


Volume vs Year
Volume
0
2
4
6
8
1990
1995
2000
2005
2010
Year


Scatterplot for Volume
Volume
Index
0
2
4
6
8
0
200
400
600
800
1000



The correlation of the data 'weekly' shows a strong correlation between the volume and the year. However, other variables have no such strong correlation. Further, the variable year and volume are visualized. From the year and volume plot, it seems like there is a gradual exponential increase from the year 1995 to 2004. For the following years, the volume increases with the year, slightly decreasing in 2010.

b. 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?

```
##
## 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
```

Based on the summary of the model, it appears the only lag2 is statistically significant with the p-value of 0.0296 at $P<0.05$. The estimated coefficient of lag2 is 0.05844 that means, when the other predictors in the model are constant, we would expect a mean increase in log odds as the stock market goes up by the unit increase in lag2. Other than this, the deviance residual of the model shows that the data is positively skewed.

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.

```
## [1] "Confusion Matrix:"
##
## preds  Down  Up
##   Down   54  48
##   Up    430 557
```

The confusion matrix revealing out correct and the wrong prediction for the model. According to this matrix, we have four different factors: True positive, True negative, False positive, and False-negative. True positive and true-negative are those which we predicted correctly. However, false positives and false negatives are those which we predicted incorrectly. In our confusion matrix, our correct prediction of the model for the direction up and down are 557 and 54 respectively. The value 48 is the false positive which means we predicted it as up but, the direction of those data was down. The value 430 is a false negative which means we predicted it as down but, the direction of those data was up. Additionally, we can also compute test error form the matrix. From the matrix `(54+556)/1089` percentage of the correct prediction is `56.10%.` We also can say that the if the model goes up our model will be correct at `557/48+557` 92.06%. Whereas, as the model goes down, our model will be correct at `54/54+430` i.e. 11.15%.

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.

```
##
## Call:
## glm(formula = Direction ~ Lag2, family = binomial, data = train)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.536  -1.264   1.021   1.091   1.368
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.20326    0.06428   3.162  0.00157 **
## Lag2         0.05810    0.02870   2.024  0.04298 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
```

```
##
##     Null deviance: 1354.7  on 984  degrees of freedom
## Residual deviance: 1350.5  on 983  degrees of freedom
## AIC: 1354.5
##
## Number of Fisher Scoring iterations: 4
```

```
## Down   Up
##   43   61
```

```
## [1] "Confusion Matrix:"
##
## preds  Down  Up
##   Down   32  25
##   Up    452 580
```

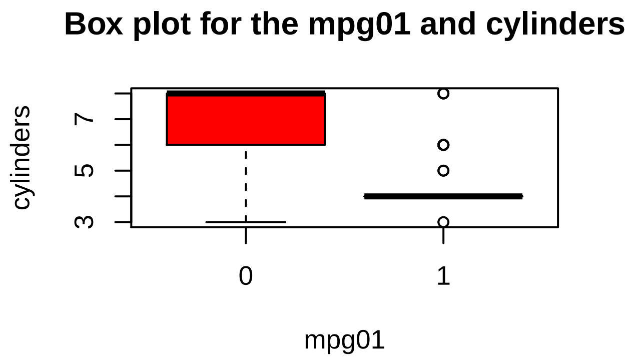
In our model, we have 43 of the total data down and 61 of the data up.In our confusion matrix, our correct prediction of the model for the direction up and down are 580 and 32 respectively. The value 25 is the false positive which means we predicted it as up but, the direction of those data was down. The value 452 is a false negative which means we predicted it as down but, the direction of those data was up. Additionally, we can also compute test error form the matrix. From the matrix `(32+580)/1089` percentage of the correct prediction is `56.19%.` We also can say that the if the model goes up our model will be correct at `580/25+580` 95.86%. Whereas, as the model goes down, our model will be correct at `32/32+580` i.e. 5.22%.

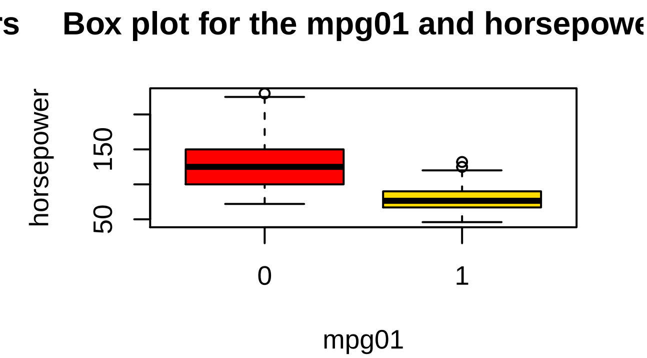
3. Question 4.7.11(a,b,c,f) pg 172
4. In this problem, you will develop a model to predict whether a given car gets high or low gas mileage based on the Auto data set.

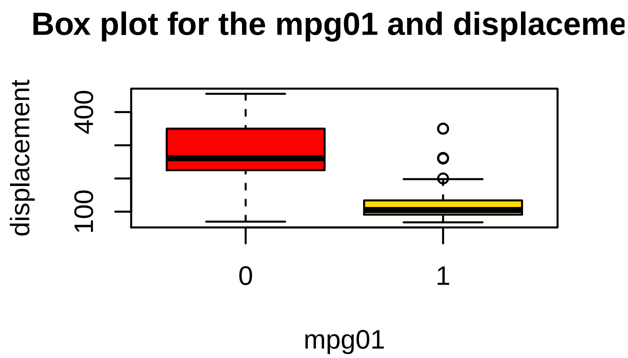
a. Create a binary variable, mpg01, that contains a 1 if mpg contains a value above its median, and a 0 if mpg contains a value below its median. You can compute the median using the median() function. Note you may find it helpful to use the data.frame() function to create a single data set containing both mpg01 and the other Auto variables.

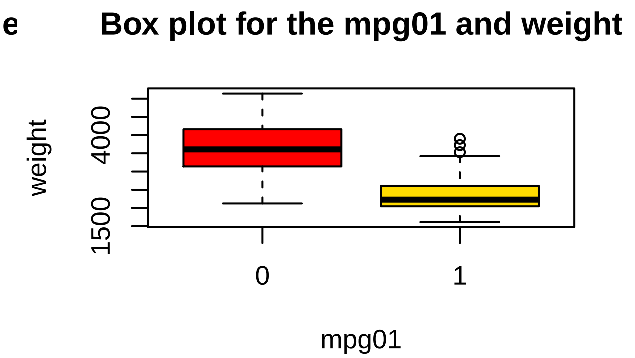
```
##   mpg cylinders displacement horsepower weight acceleration year origin
## 1  18         8          307        130   3504         12.0   70      1
## 2  15         8          350        165   3693         11.5   70      1
## 3  18         8          318        150   3436         11.0   70      1
## 4  16         8          304        150   3433         12.0   70      1
## 5  17         8          302        140   3449         10.5   70      1
## 6  15         8          429        198   4341         10.0   70      1
##                        name mpg01
## 1 chevrolet chevelle malibu     0
## 2         buick skylark 320     0
## 3        plymouth satellite     0
## 4             amc rebel sst     0
## 5               ford torino     0
## 6          ford galaxie 500     0
```

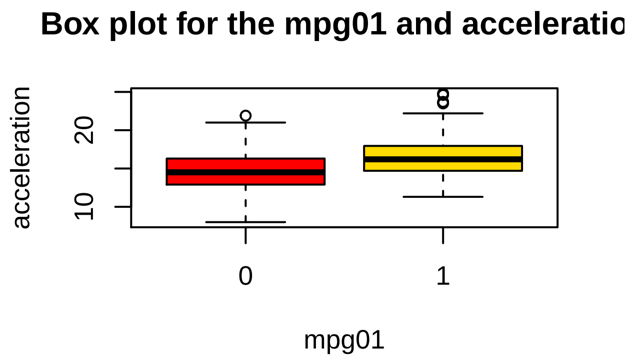
b) Explore the data graphically in order to investigate the association between mpg01 and the other features. Which of the other features seem most likely to be useful in predicting mpg01? Scatterplots and boxplots may be useful tools to answer this question. Describe your findings.

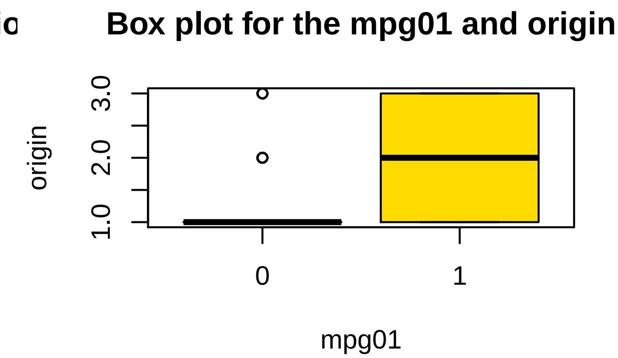

Box plot for the mpg01 and cylinders
cylinders
3
5
7
0
1
mpg01

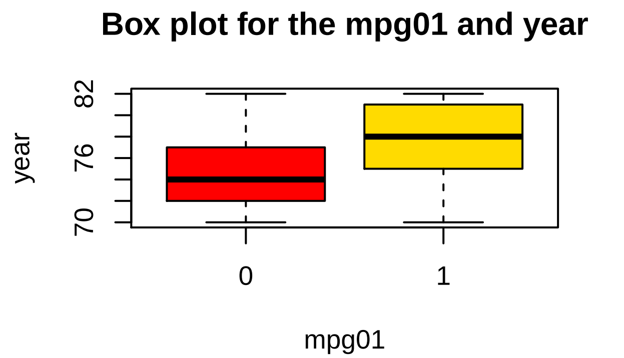

Box plot for the mpg01 and horsepowe
horsepower
50
150
0
1
mpg01

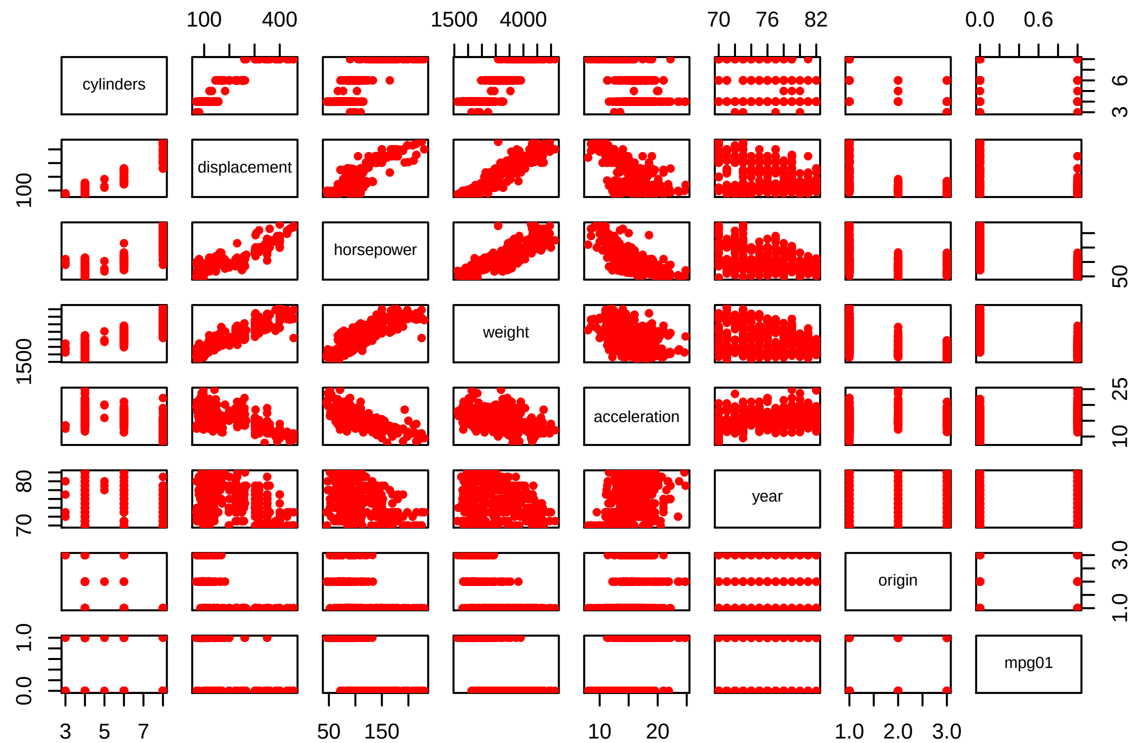

Box plot for the mpg01 and displaceme
displacement
100
400
0
1
mpg01


Box plot for the mpg01 and weight
weight
1500
4000
0
1
mpg01

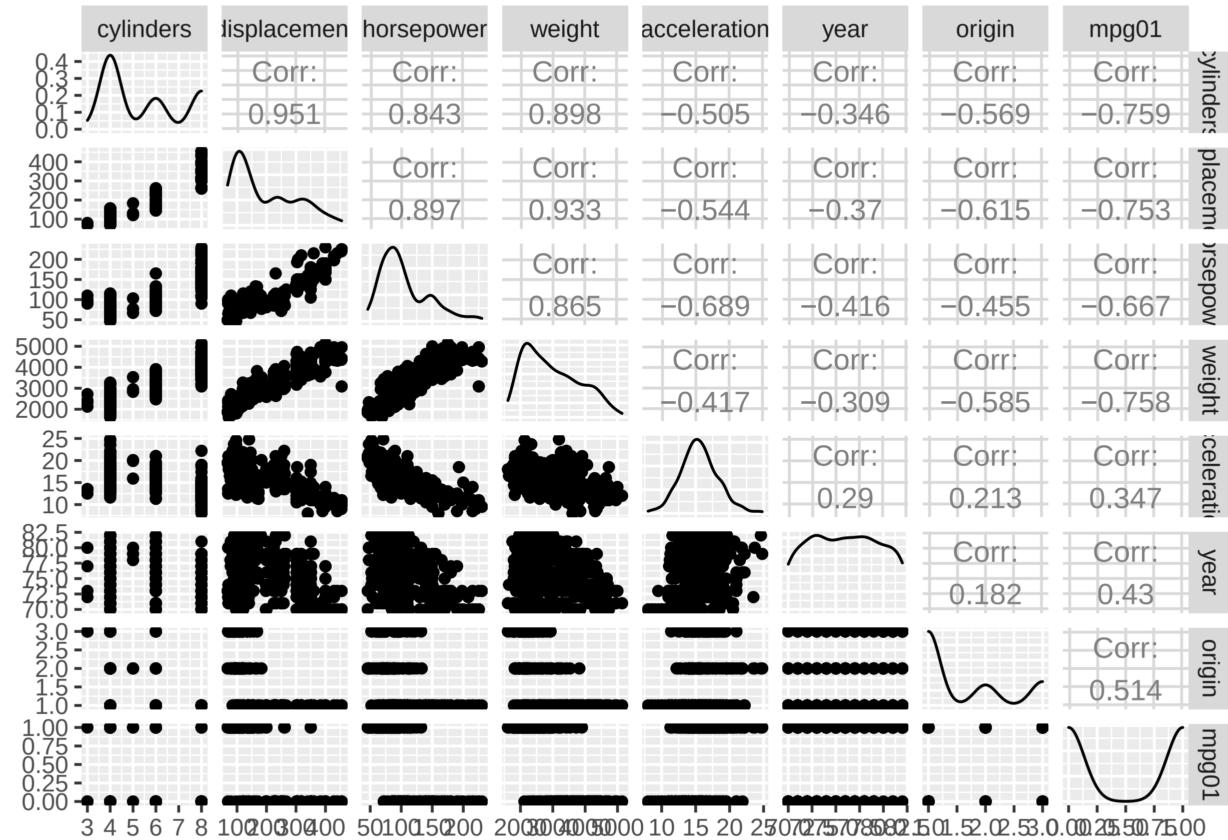

Box plot for the mpg01 and acceleratio
acceleration
10
20
0
1
mpg01


Box plot for the mpg01 and origin
origin
1.0
2.0
3.0
0
1
mpg01


Box plot for the mpg01 and year
year
70
76
82
0
1
mpg01




```
##              cylinders displacement horsepower     weight acceleration
## cylinders    1.0000000    0.9508233  0.8429834  0.8975273   -0.5046834
## displacement 0.9508233    1.0000000  0.8972570  0.9329944   -0.5438005
## horsepower   0.8429834    0.8972570  1.0000000  0.8645377   -0.6891955
## weight       0.8975273    0.9329944  0.8645377  1.0000000   -0.4168392
## acceleration -0.5046834  -0.5438005 -0.6891955 -0.4168392    1.0000000
## year         -0.3456474  -0.3698552 -0.4163615 -0.3091199    0.2903161
## origin       -0.5689316  -0.6145351 -0.4551715 -0.5850054    0.2127458
## mpg01        -0.7591939  -0.7534766 -0.6670526 -0.7577566    0.3468215
##                   year     origin      mpg01
## cylinders    -0.3456474 -0.5689316 -0.7591939
## displacement -0.3698552 -0.6145351 -0.7534766
## horsepower   -0.4163615 -0.4551715 -0.6670526
## weight       -0.3091199 -0.5850054 -0.7577566
## acceleration  0.2903161  0.2127458  0.3468215
## year          1.0000000  0.1815277  0.4299042
## origin        0.1815277  1.0000000  0.5136984
## mpg01         0.4299042  0.5136984  1.0000000
```



From the box plot, it is clear that there is a clear distinction between the distribution in two groups for the variables cylinders, horsepower, displacement weight, origin, and year. We also can notice that most of the automobiles were originated in Japan. US-based cars are mostly condensed at lower mpg, whereas European and Japanese cars tend to be well distributed. Also, older cars tend to have lower mpg, and modern cars tend to have higher.Also, older cars tend to have lower mpg, and modern cars tend to have higher. From the correlation plot, it looks like the physical quantities of the car are highly correlated.The displacement and the horsepower look to have an exponential relationship.

c. Split the data into a training set and a test set.

   We splitted the data in the ration of 70% and 30% .

d. 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?

```
##
## Call:
## glm(formula = mpg01 ~ cylinders + weight + displacement + horsepower,
##     family = binomial, data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4956  -0.1154   0.0728   0.2892   1.9696
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)  11.6726983  2.1736488   5.370 7.87e-08 ***
## cylinders     0.7982266  0.4492415   1.777  0.07560 .
```

```
## weight       -0.0021338  0.0008688  -2.456  0.01405 *
## displacement -0.0291275  0.0112576  -2.587  0.00967 **
## horsepower   -0.0491982  0.0174464  -2.820  0.00480 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 381.23  on 274  degrees of freedom
## Residual deviance: 131.96  on 270  degrees of freedom
## AIC: 141.96
##
## Number of Fisher Scoring iterations: 7

##
## preds  0  1
##     0 49  5
##     1 10 53

## [1] "Test error (percantage): 12.82"
```

From question b, we had found that cylinders, weight, displacement, horsepower were mostly associated with the variable mpg01. Hence, we have performed logistic regression with these variables. For the computed model, we found out that the weight and the horsepower are statistically significant. Also, the data is of the model is negatively skewed. For the test accuracy, I have computed the confusion matrix and then found the accuracy of the model and the test error. The confusion matrix shows that we were able to predict 88.14% of the data correctly. Likewise, we predicted 11.86 % of the data incorrectly.Therefore, we have 11.86% as the test error.

4. Write a reusable function in RMD that calculates the misclassification rate, sensitivity, and specificity, and return a table similar to `Table 4.7`. Call this function `misclass.fun.*`, replacing `*` with your initials. The arguments for this function are a threshold, predicted probabilities, and original binary response data. Test your function using the data and model from 4.7.10 b) with threshold values of `c(0.25, 0.5, 0.75)`.

Post any questions you might have regarding this on the discussion board. Define `misclass.fun.*` using the `function()` command. Open code that is not using `function()` will not be graded. We will calculate misclassification rates frequently this semester, so take care that you write a reusable function in order to save time this semester. *Show the function code you wrote in your final write-up using* `echo = T`.

```r
# thd <- 0.75
misclass.fun.yd <- function(thd, pred_prob, original_res){
  predicted=rep("Down",length(original_res))
  vals <- pred_prob
  for(i in 1:length(original_res)){
    if(vals[i]>=thd){
      predicted[i]="Up"
    }
  }
   con.mat = table(predicted, original_res) # creating confusion matrix
    # since all the pred values for threshold 0.25 are less than 0.25 therefore we only
   # have 1 row as the confusion matrix therefore checking the row
   if(length(con.mat)==2){
      MCR = mean(predicted != original_res) #misclassification rate
      SEN = con.mat[1, 2] / sum(con.mat[1,]) # sensitivity
      SPEC = con.mat[1, 1] / sum(con.mat[1,]) # specificity
    }else{
```

```
      MCR = (con.mat[1, 2] + con.mat[2, 1]) / sum(con.mat) # misclassification rate
      SEN = con.mat[2, 2] / (con.mat[2, 2] + con.mat[1, 2]) # sensitivity
      SPEC = con.mat[1, 1] / (con.mat[1, 1] + con.mat[2, 1]) # specificity
      }


   return(list(
    Misclassification_Rate = MCR,
    Sensitivity = SEN,
    Specificity = SPEC
  ))

}

pred_prob<- predict(fit_log, newdata = Weekly, type ="response") # model form the q 4.7.10(b)
original_res <- Weekly$Direction

at_0.25threshold <- misclass.fun.yd(0.25,pred_prob,original_res)
at_0.5threshold <- misclass.fun.yd(0.5,pred_prob,original_res)
at_0.75threshold <- misclass.fun.yd(0.75,pred_prob,original_res)


library(knitr)
finaltable <- as.data.frame(cbind(at_0.25threshold, at_0.5threshold, at_0.75threshold))
knitr::kable(finaltable, digits = 3,
             caption = "Different measure of accuracy with different threshold")
```

Table 1: Different measure of accuracy with different threshold

	at_0.25threshold	at_0.5threshold	at_0.75threshold
Misclassification_Rate	0.4444444	0.4389348	0.5546373
Sensitivity	0.5555556	0.9206612	0.003305785
Specificity	0.4444444	0.1115702	0.9979339