Assignment 3
Shaheed Shihan
Modern Applied Statistics II

Chapter 4

1. 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.

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

10. 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.

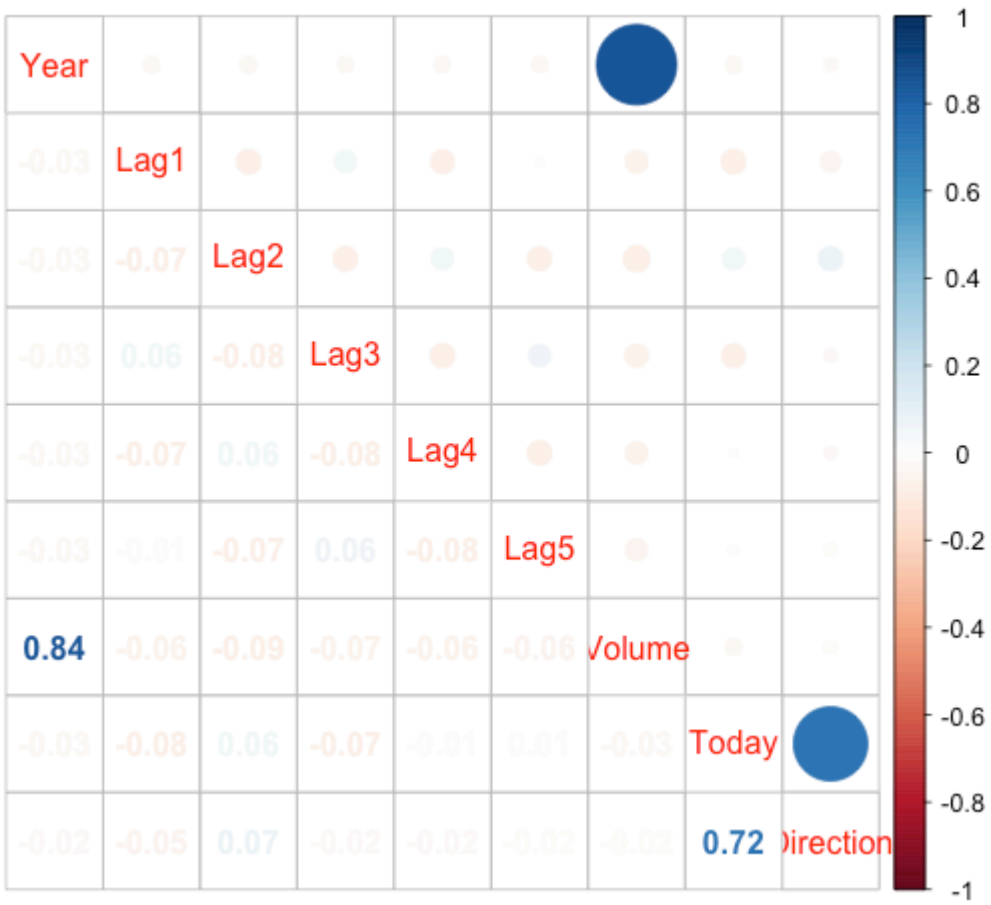
(a) Produce some numerical and graphical summaries of the Weekly data. Do there appear to be any patterns?

```
> summary(Weekly)
```

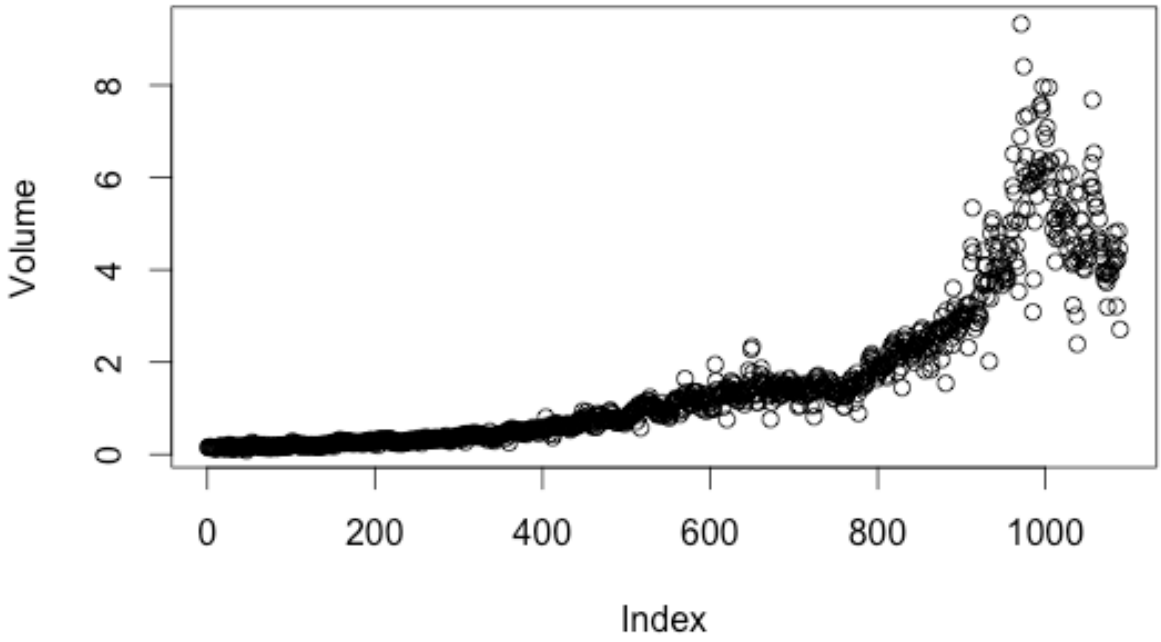
Year	Lag1	Lag2	Lag3	Lag4
Min. :1990	Min. :-18.1950	Min. :-18.1950	Min. :-18.1950	Min. :-18.1950
1st Qu.:1995	1st Qu.: -1.1540	1st Qu.: -1.1540	1st Qu.: -1.1580	1st Qu.: -1.1580
Median :2000	Median : 0.2410	Median : 0.2410	Median : 0.2410	Median : 0.2380
Mean :2000	Mean : 0.1506	Mean : 0.1511	Mean : 0.1472	Mean : 0.1458
3rd Qu.:2005	3rd Qu.: 1.4050	3rd Qu.: 1.4090	3rd Qu.: 1.4090	3rd Qu.: 1.4090
Max. :2010	Max. : 12.0260	Max. : 12.0260	Max. : 12.0260	Max. : 12.0260

Lag5	Volume	Today	Direction
Min. :-18.1950	Min. :0.08747	Min. :-18.1950	Down:484
1st Qu.: -1.1660	1st Qu.:0.33202	1st Qu.: -1.1540	Up :605
Median : 0.2340	Median :1.00268	Median : 0.2410	
Mean : 0.1399	Mean :1.57462	Mean : 0.1499	
3rd Qu.: 1.4050	3rd Qu.:2.05373	3rd Qu.: 1.4050	
Max. : 12.0260	Max. :9.32821	Max. : 12.0260	

Correlation plot:



Volume and Year seemed to be highly correlated. This was further investigated using a simple plot.



(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?

```
> summary(glm.fit)

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
```

Besides the intercept, only lag 2 is somewhat significant. Since the p-value is 0.02, which is relatively large, there is no clear evidence of a real association between lag2 and Direction.

(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.

```
> table(glm.pred, Direction)
```

	Direction	
glm.pred	Down	Up
Down	54	48
Up	430	557

```
> (54+557)/1089
[1] 0.5610652

> mean(glm.pred==Direction)
[1] 0.5610652
```

The diagonal elements of the confusion matrix indicate correct prediction, while the off-diagonals represent the incorrect ones. We had a total of 571 (54+557) correct predictions and 478 (430+48) incorrect predictions. That gave us an accuracy rate of 56.1%.

(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).

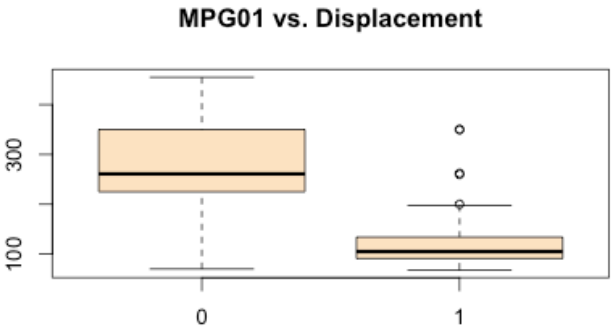
```
> table(glm.pred2, Direction.2008)
                Direction.2008
```

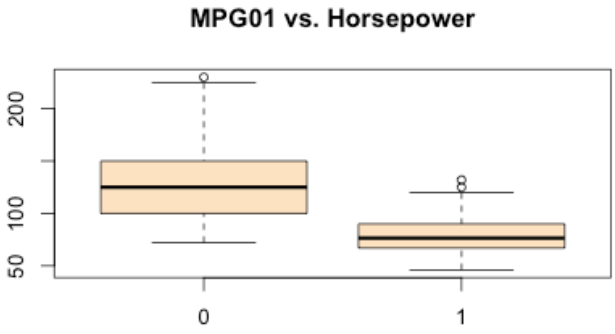
```
glm.pred2      Down  Up
   Down        9     5
   Up          34    56
> mean(glm.pred2 == Direction.2008)
[1] 0.625
> (56+9)/104
[1] 0.625
```

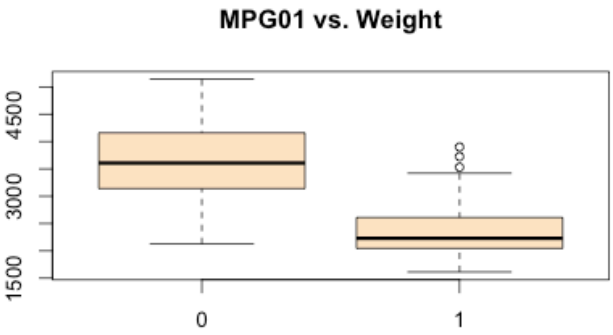
The results have improved. Using only lag2 as our predictor variable, we predicted the direction right almost 63% of the time.

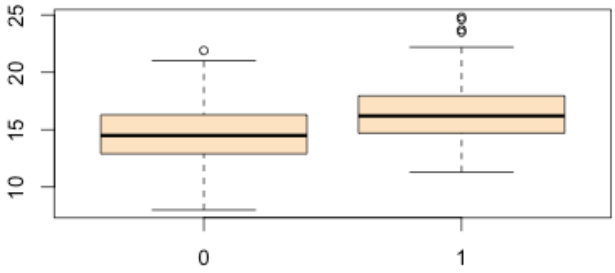
11. In this problem, you will develop a model to predict whether a given car gets high or low gas mileage based on the Autodata set.
(a) Create a binary variable, mpg01 that contains a 1 if mpg contains a value above its median, and a 0 if mpg contains a value below its median. You can compute the median using the median () function. Note you may find it helpful to use the data.frame() function to create a single data set containing both mpg01and the other Auto variables.

(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?



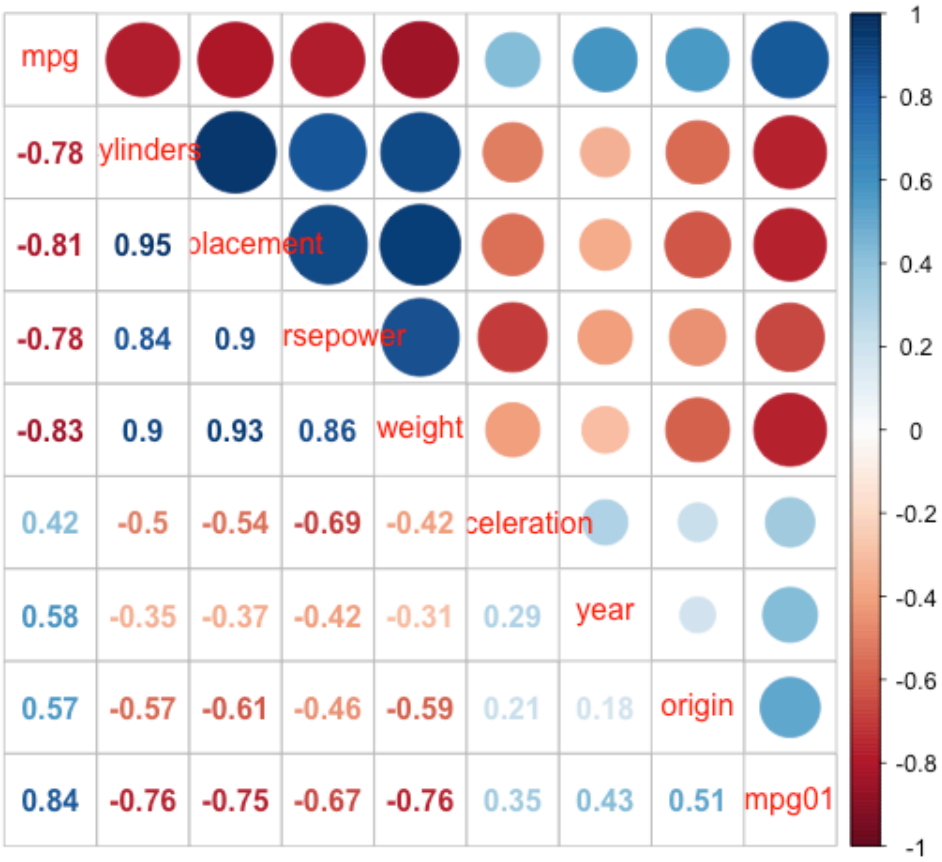






Scatterplots and boxplots may be useful tools to answer this question. Describe your findings.

MPG01 seemed closely correlated to displacement, horsepower, acceleration and cylinders. However, at the same time, there exists Multicollinearity between the predictor variablesas shown by the correlation plot. Further investigation into the data set is necessary.



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

See R code.

(e) 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?

glm.pred3	0	1
0	45	3

```
     1             4 46
> mean(glm.pred3 == A.test$mpg01)
[1] 0.9285714
> (45+46)/98
[1] 0.9285714
```

We had a prediction rate of 92.8% which was surprisingly high. Switching around some of the predictor variables had little effect on the prediction rate. Including more variables actually increased the prediction rate.

4. **See R code.**