Exercise 13, p. 193 (a), (b), (c), (d) only: logistic regression and prediction using the Weekly data set.

```
library(ISLR)
weekly = read.csv("/Volumes/work/MTH522/data/Weekly.csv")
head(weekly)
```

```
##   Year   Lag1   Lag2   Lag3   Lag4   Lag5    Volume  Today Direction
## 1 1990  0.816  1.572 -3.936 -0.229 -3.484 0.1549760 -0.270      Down
## 2 1990 -0.270  0.816  1.572 -3.936 -0.229 0.1485740 -2.576      Down
## 3 1990 -2.576 -0.270  0.816  1.572 -3.936 0.1598375  3.514        Up
## 4 1990  3.514 -2.576 -0.270  0.816  1.572 0.1616300  0.712        Up
## 5 1990  0.712  3.514 -2.576 -0.270  0.816 0.1537280  1.178        Up
## 6 1990  1.178  0.712  3.514 -2.576 -0.270 0.1544440 -1.372      Down
```

```
head(weekly)
```

```
##   Year   Lag1   Lag2   Lag3   Lag4   Lag5    Volume  Today Direction
## 1 1990  0.816  1.572 -3.936 -0.229 -3.484 0.1549760 -0.270      Down
## 2 1990 -0.270  0.816  1.572 -3.936 -0.229 0.1485740 -2.576      Down
## 3 1990 -2.576 -0.270  0.816  1.572 -3.936 0.1598375  3.514        Up
## 4 1990  3.514 -2.576 -0.270  0.816  1.572 0.1616300  0.712        Up
## 5 1990  0.712  3.514 -2.576 -0.270  0.816 0.1537280  1.178        Up
## 6 1990  1.178  0.712  3.514 -2.576 -0.270 0.1544440 -1.372      Down
```

```
dim(weekly)
```

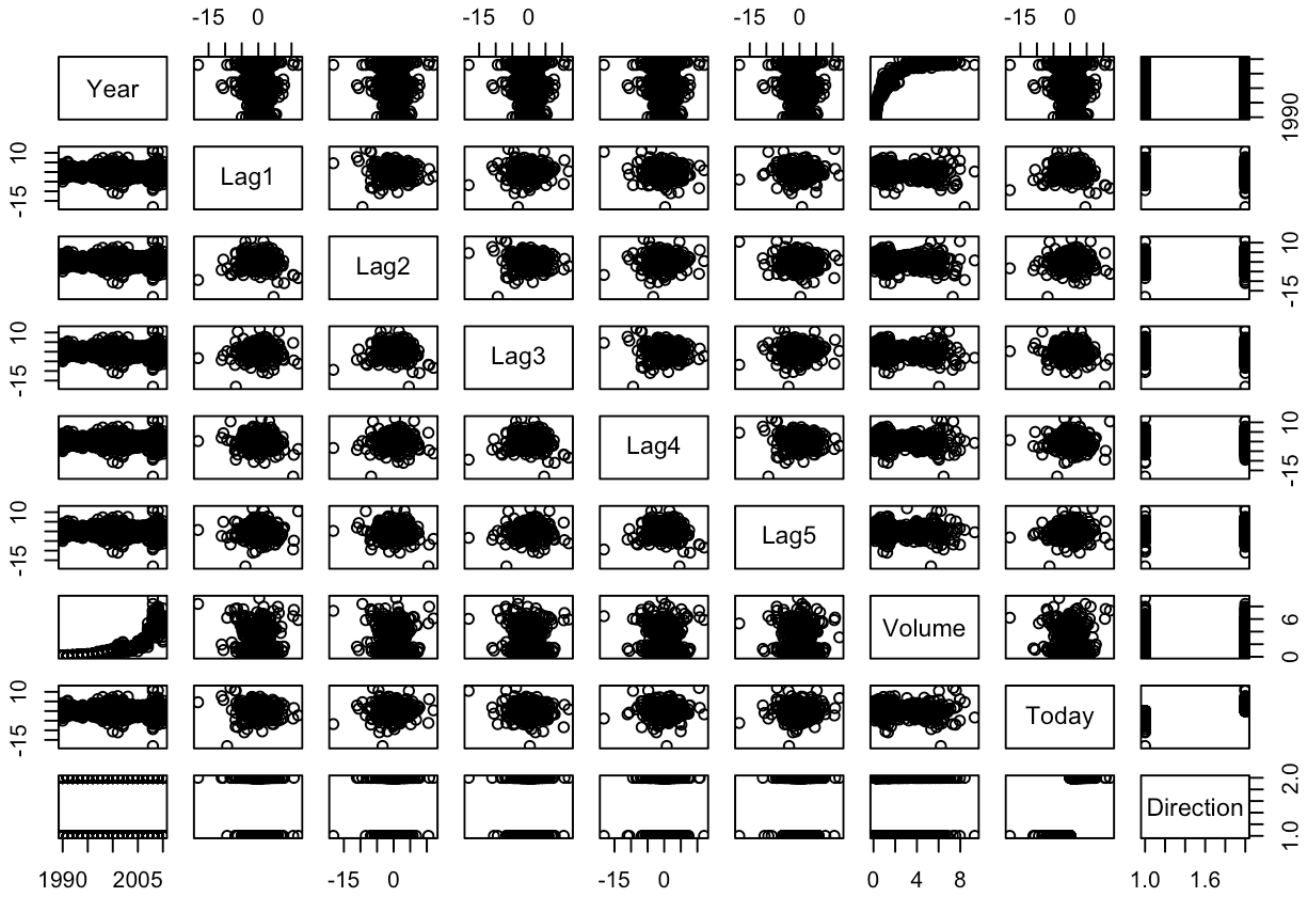
```
## [1] 1089    9
```

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

```
summary(weekly)
```

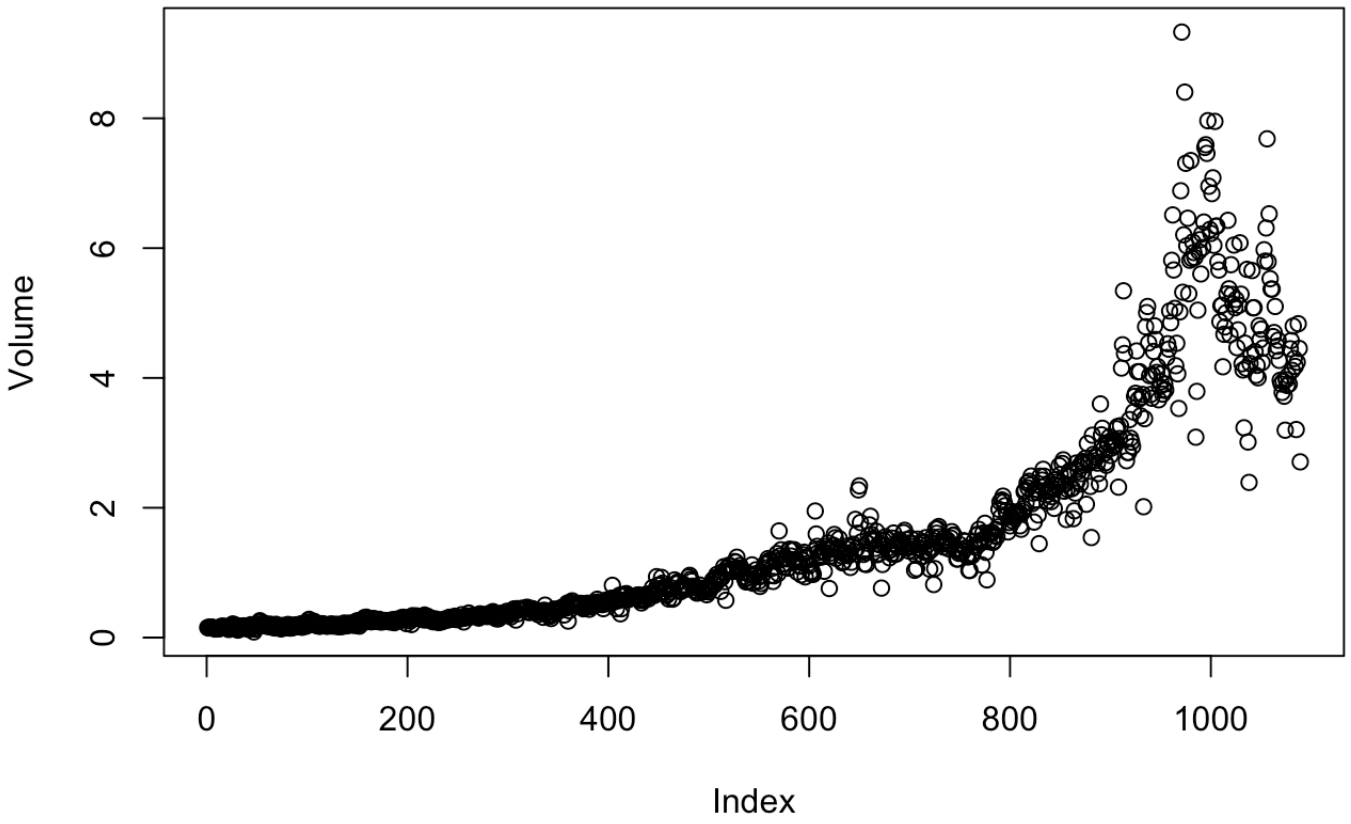
```
##       Year           Lag1                Lag2                Lag3          
##  Min.   :1990   Min.   :-18.1950   Min.   :-18.1950   Min.   :-18.1950  
##  1st Qu.:1995   1st Qu.: -1.1540   1st Qu.: -1.1540   1st Qu.: -1.1580  
##  Median :2000   Median :  0.2410   Median :  0.2410   Median :  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        
##  Length:1089        
##  Class :character   
##  Mode  :character   
##                     
##                     
##
```

```
pairs(Weekly)
```



Observations: 1. There isn't much analysis we can do here, only thing we can say is that, `volume` of shares increased throught the period i.e. from 1990 to 2010.

```
attach(weekly)
plot(Volume,)
```



Observations: 1. Looking at scatterplot of volume over time, we can see that the number of shares traded each week has grown exponentially over the years from 1990 to 2010 in the data.

```
cor(Weekly[-9])
```

```
##                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
```

Observations: 1. We can see that each of the lag variables is only correlated very weakly with today's returns. 2.The sole substantial value of 0.842, between Volume and Year, aligns with the strong correlation we saw in the above scatterplot.

13(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?

```
model_glm = glm(Direction ~ . - Year - Today, data = Weekly, family = "binomial")
summary(model_glm)
```

```
##
## Call:
## glm(formula = Direction ~ . - Year - Today, 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
```

Observations: 1. From the above summary, `Lag2` has the smallest p-value and is the only one close to zero with a value of `0.0296`, providing evidence at the 5% significance level to reject the null hypothesis that it is not related to the response `Direction`. 2. `Lag1` is somewhat near the border of being significant at the 10% level, with a p-value of `0.1181`. 3. None of the above predictors are statistically significant.

13(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.

```
model_prob = predict(model_glm,type = "response")
model_predict <- rep("Down",1089)
model_predict[model_prob > .5] = "Up"
table(model_predict, Direction)
```

```
##               Direction
## model_predict Down  Up
##          Down   54  48
##          Up    430 557
```

```
mean(model_predict == Direction)
```

```
## [1] 0.5610652
```

Observations: 1. Our model correctly predicted 54 down weeks out of a total of 484 actual down weeks and 557 up days out of a total of 605 actual up weeks. This means that the model correctly predicted the direction for 611 weeks out of the 1089 for an accuracy of 0.5612. 2. The true positive rate is the number of correctly predicted positives divided by the overall number of positives. So for this model -> 557/605 ≈ 0.92. 3. The false positive rate is the number of incorrectly predicted positives (weeks incorrectly predicted to be up weeks = 430 weeks) divided by the overall number of negatives (the total number of down weeks = 484 weeks) – is comparably high at 430/484≈0.888. 4. The positive predictive value, which is the number of true positives divided by the total number of predicted positives, so 557/987 ≈ 0.564. 5. The negative predictive value, which is the number of true negatives divided by the total number of predicted negatives; so 54/102 ≈ 0.529.

13(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).

```
train <- (Year < 2009)
weekly.2009 <- weekly[!train, ]
Direction.2009 <- Direction[!train]
```

```
model_fit = glm(Direction ~ Lag2, data = Weekly, subset = train, family = "binomial")
summary(model_fit)
```

```
##
## Call:
## glm(formula = Direction ~ Lag2, family = "binomial", data = Weekly,
##     subset = 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
```

```
model_prob_2 <- predict(model_fit, weekly.2009, type = "response")
model_predict_2 <- rep("Down", 104)
model_predict_2[model_prob_2 > .5] <- "Up"
table(model_predict_2, Direction.2009)
```

```
##                Direction.2009
## model_predict_2 Down Up
##            Down    9  5
##            Up     34 56
```

```
mean(model_predict_2 == Direction.2009)
```

```
## [1] 0.625
```

```
mean(model_predict_2 != Direction.2009)
```

```
## [1] 0.375
```

Observations: 1. After fitting a logistic regression model on the data with only `Lag2` as the predictor, the model correctly predicted the market direction for 62.5% of the weeks in the held-out data (the data from 2009 and 2010). 2. Continuing with the convention from Part 3 that an up week is a positive result, the true positive rate is 56/61≈0.918, and false positive rate is 34/43 ≈ 0.791. The positive predictive value is 56/90 ≈ 0.622 and the negative predictive value is 9/14 ≈ 0.643.