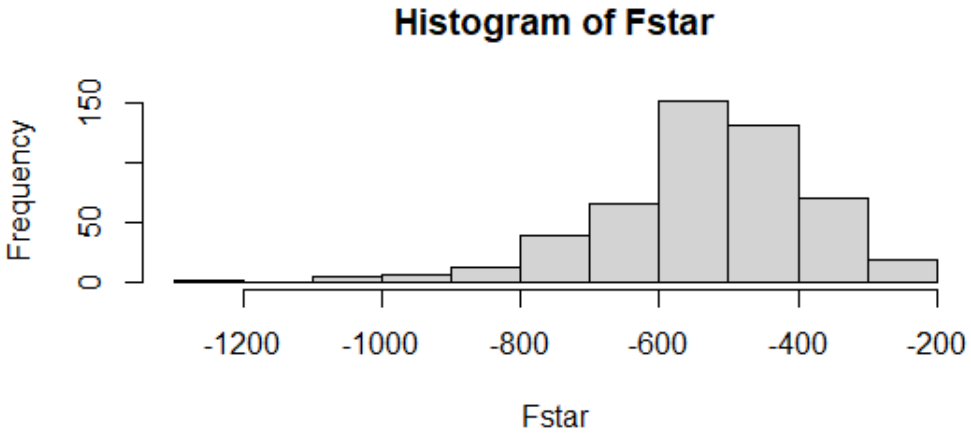
Neha Maddali

Problem 1:

a. Confidence interval (95%) = (-0.4512375, -0.2181824)



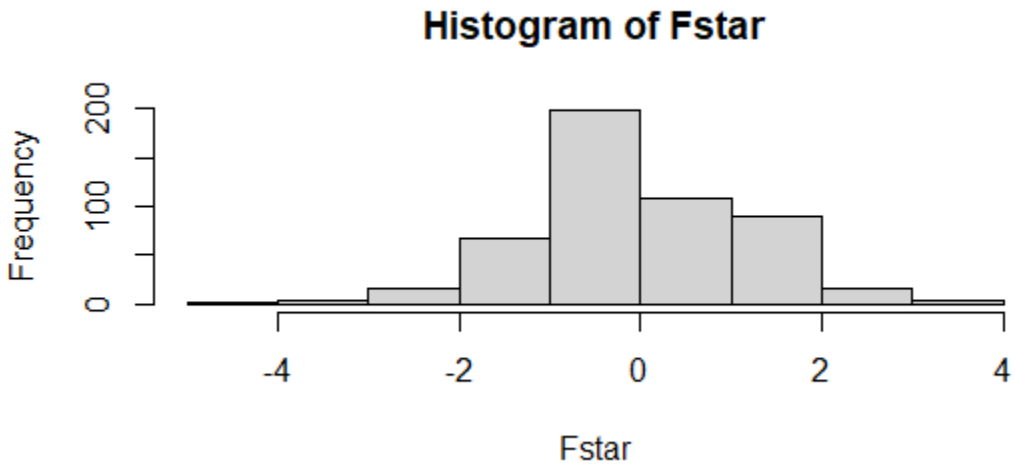
b. Analytic confidence interval is (-0.4004172, -0.22321431). This is very close to the confidence interval that I found with bootstrap.

c. $\mu\hat{}_{med} = 21.2$

d.
```
B = 2000
median_boot = rep(NA,2000)
for(b in 1:B){
  index = sample(1:n,n,replace=TRUE)
  bootstrap = Boston[index,]
  median_boot[b] = median(bootstrap$medv, na.rm=TRUE)
}
sqrt(sum((median_boot-mean(median_boot))^2)/(B-1))
```

standard error = 0.374

e. confidence interval: (20.30754, 21.97214)



f. $\mu\hat{}_{0.1} = 12.8$

g. standard error = 0.496. This shows how far the estimate of the 10th percentile of medv is from the true value

Problem 2:

a. Spam proportion: 1813 / 4601 = 39.4%
   Non-spam proportion: 2788 / 4601 = 60.6%

b. Training spam proportion: 927 / 2300 = 40.3%
   Training non-spam proportion: 1373 / 2300 = 59.7%
   Testing spam proportion: 886 / 2301 = 38.5%
   Testing non-spam proportion: 1415 / 2301 = 61.5%
   With a 50-50 split for the training and test set, the proportions are relatively the same from part a.

c. First 10 predicted probabilities:

```
        1         2         3         5         6         8
0.6830159 0.9962195 0.9999999 0.7700588 0.6334144 0.6216763
       10        12        13        18
0.9259383 0.5188221 0.6779338 0.9998643
```

d.

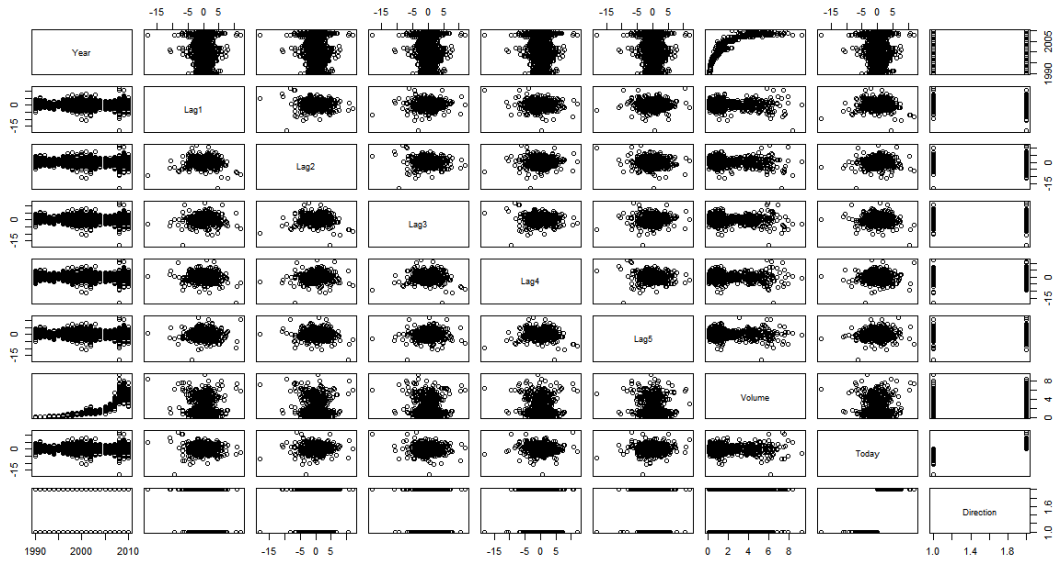
```
preds    0    1
    0 1351   89
    1   64  797
```

   The model predicted the spam trend correctly 93.35% of the time.
   Misclassification rate = 0.06649283
   False negative rate: 0.06180556
   False positive rate: 0.07433217

e. I think reporting meaningful email as spam is a more critical mistake. To accommodate this, we could increase the 0.5 threshold for classifying an email as spam so its harder to classify an email as spam.

Problem 3:

a. Year and Volume appear to have a logistic relationship.



b. Lag1, Lag3, Lag4, Lag5, and Volume seem to be statistically significant.

```
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
```

c.

```
preds  Down  Up
  Down   54  48
  Up    430 557
```

To determine the % of current predictions: (54+557) / (54+48+430+557) = 0.5611
This says that the model predicted the weekly market trend correctly 56.11% of the time. Separating in how the model correctly predicts the Up and Down trends. The model correctly predicted the Up weekly trends (557) / (48+557) = 0.9207 which is 92.07% correct. The Down weekly trends were predicted at a lower rate, (54) / (430 + 54) = 0.115 which is 11.5% correctly predicted.

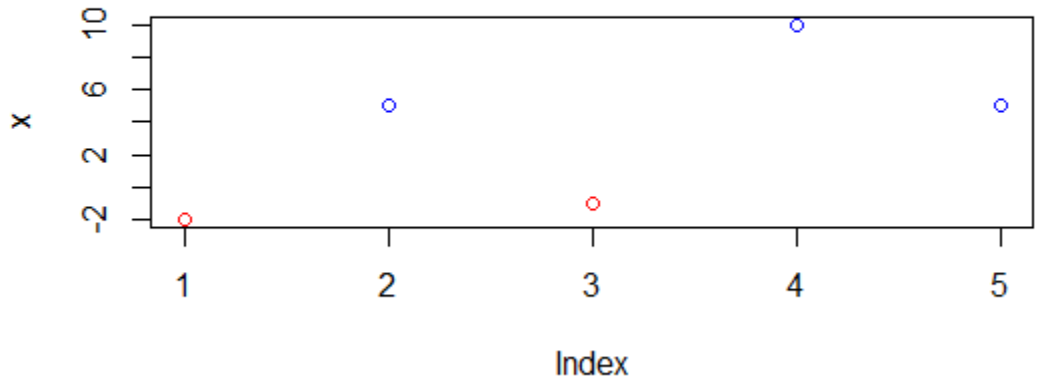
d.

```
              Direction.train
logWeekly.pred Down Up
          Down    9  5
          Up     34 56
```

When splitting the Weekly dataset into training and test data, the model correctly predicted weekly trends at a rate of 62.5%, which is an improvement from the model that used the whole dataset. This model predicted upward trends as 91.8% and downwards as 20.93% correct. This model was able to improve significantly on correctly predicting downwards trends. The overall fraction of correct predictions is 62.5%

Problem 4:

a.



The two groups are separated. The red data points are negative values while the blue data points are positive values.

b. The following error is printed::
   Error in eval(family$initialize) : y values must be 0 <= y <= 1
c. B0 = infinity, B1 = infinity
d. It looks like the main limitation is that there is an assumption of linearity between the dependent and independent variables. There is no closed form analytical solution to it.