Assignment 2

By Mrunal Ghorpade

UIN: 677441117 /Net ID: mghorp2

1) Show that the ridge regression estimate is the mean (and mode) of the posterior distribution, under a Gaussian prior $\beta \sim \mathbb{N}(0, \tau I)$, and Gaussian sampling model $y \sim \mathbb{N}(X\beta, \sigma 2I)$. Find the relationship between the regularization parameter λ in the ridge formula, and the variances τ and $\sigma 2$.

Ans:

Gaussian sampling model $y \sim N(X\beta, \sigma^2 I)$
and gaussian prior $\beta \sim N(0, \tau I)$

Posterior distribution is given as;

$$P(\theta|Data) = \frac{P(Data|\theta)\ P(\theta)}{P(Data)}$$

Taking log of above

$$\log P(\theta|Data) = \log P(Data|\theta) + \log P(\theta) - \log P(Data) \quad \text{---(1)}$$

$$\cancel{\log} P(Data|\theta) = P(Y=y_1, X=x_1)\, P(Y=y_2, X=x_2) \cdots$$
$$= \sum_{i=1}^{N} P(Y=y_i \mid X=x_i)$$

$$\log P(Data|\theta) = \log \sum_{i=1}^{N} P(Y=y_i \mid X=x_i)$$

Assum x is not random.
& $Y = x^T\beta + \varepsilon$ where $\varepsilon \sim N(0, \sigma^2)$
as x is fixed β has normal distribution with mean 0 & variance τ

$$\therefore \log P(Data|\theta) = \log P(Y=y, X=x)$$

$$\log P(\theta) = \log P(\beta)$$

Now using the expression of gaussian distribution

$$\log P(Data|\theta) = \log P(Y=y, X=x)$$
$$= \log\left(\frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(y_i - x_i^T\beta)^2}{2\sigma^2}\right\}\right)$$

and $$\log P(\theta) = \log\left(\frac{1}{\sqrt{2\pi\tau}} \exp\left\{-\frac{\beta^2}{2\tau}\right\}\right)$$

Substituting above in Eq^n (1) we get;

$$\log P(\Theta \mid Data) = \log\left(\frac{1}{\sqrt{2\pi}\sigma} \exp\left\{\frac{-(y_i - x_i^T \beta)^2}{2\sigma^2}\right\}\right) + \log\left(\frac{1}{\sqrt{2\pi\tau}} \exp\left\{\frac{-\beta^2}{2\tau}\right\}\right) - \log P(Data)$$

Taking derivative w.r.t β

$$\frac{d \log P(\Theta \mid Data)}{d\beta} = \frac{d}{d\beta}\left\{\log \frac{1}{\sqrt{2\pi}\sigma} - \frac{(y_i - x_i^T\beta)^2}{2\sigma^2} + \log \frac{1}{\sqrt{2\pi\tau}} - \frac{\beta^2}{2\tau} - \log P(Data)\right\}$$

$$= \frac{d}{d\beta}\left(\frac{1}{\sigma^2}(y_i - x_i^T\beta)^2 - \frac{2\beta}{\tau}\right)$$

$$= \left(-\frac{2X^TY}{\sigma^2} + \frac{2X^TX\beta}{\sigma^2}\right) - \frac{2\beta}{\tau}$$

Setting the above Eq^n to 0

$$0 = X^TY - X^TX\beta - \frac{\sigma^2}{\tau}\beta$$

$$\therefore \beta = \left(X^TX + \frac{\sigma^2}{\tau} I\right)^{-1} X^TY \qquad -(2)$$

Now the $\hat{\beta}$ for Ridge regression is given by

$$\hat{\beta} = (X^TX + \lambda I)^{-1} X^TY \qquad -(3)$$

$\therefore$ from Eq^n (2) & (3) we can say tha $\boxed{\lambda = \frac{\sigma^2}{\tau}}$

$\therefore$ It is clear that $P(\beta \mid Y)$ is Gaussian and its mean & median coincide.

2) This problem involves the Boston data set, which we saw in the lab for this chapter. We will now try to predict per capita crime rate using the other variables in this data set. In other words, per capita crime rate is the response, and the other variables are the predictors.

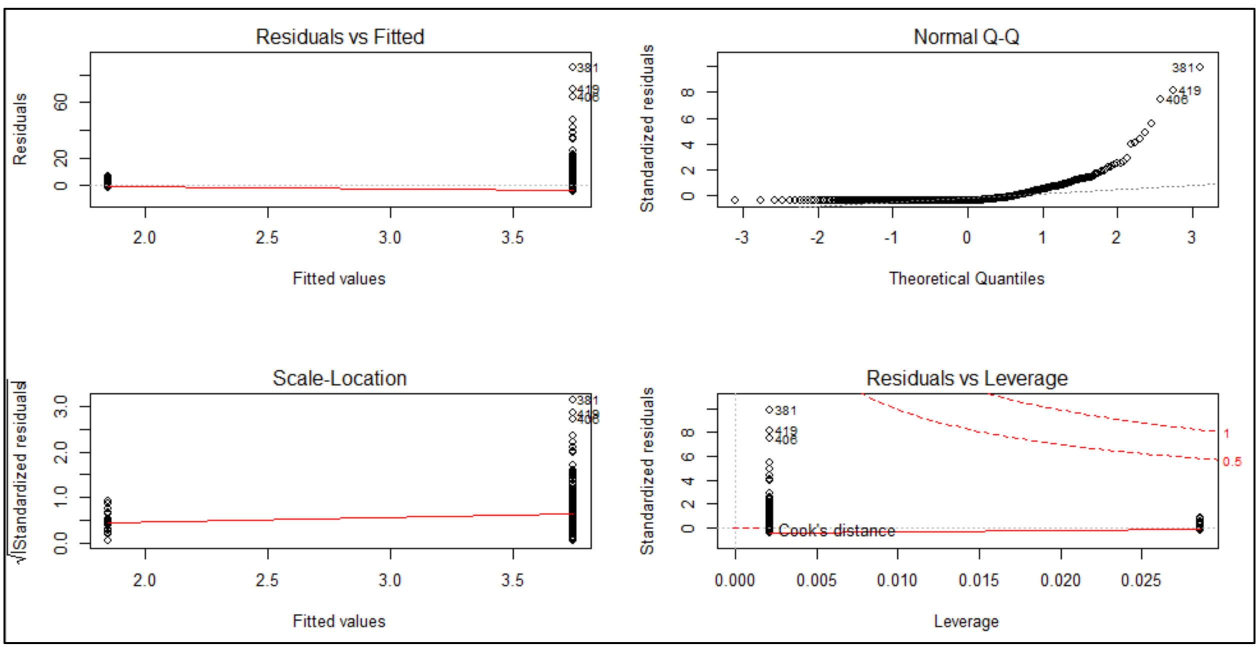
(a) For each predictor, fit a simple linear regression model to predict the response. Describe your results. In which of the models is there a statistically significant association between the predictor and the response? Create some plots to back up your assertions.

Assuming alpha to be 5% and p-values of all the predictor variables except “chas” are less than 5% so except for predictor “chas” there is statistically significant relationship between each of the predictor with response variable “crime”.

```
print(summ)
 Predictor              Estimate            Std_error             t_value               p_value
        zn  -0.073934977404123   0.0160945961932254 -4.59377647730262 5.50647210767964e-06
     indus   0.509776331104228   0.0510243323317304  9.99084765656433 1.45034893302756e-21
      chas   -1.89277655080378     1.50611548365537 -1.25672737007522     0.209434501535197
       nox    31.2485312011229     2.99919038061173  10.4189888721733  3.7517392603569e-23
        rm   -2.68405122411395    0.532041083377015 -5.04481948476145 6.34670298468749e-07
       age   0.107786227139533   0.0127364362587918  8.46282468262108 2.85486935024409e-16
       dis    -1.5509016824101    0.168330030929704 -9.21345807307415  8.5199487669261e-19
       rad   0.617910927327201   0.0343318196678424   17.998199143111 2.69384439818633e-56
       tax  0.0297422528227653  0.00184741511910325  16.0993880125884 2.35712683525685e-47
   ptratio    1.15198278707059    0.169373609252715  6.80143023552017 2.94292244735967e-11
     black -0.0362796405673308  0.00387315383170155 -9.36695058956449 2.48727397377375e-19
     lstat   0.548804782062398   0.0477609709441814   11.490653795623 2.65427723147327e-27
      medv  -0.363159922257603   0.0383901746742235 -9.45971007788721 1.17398708219449e-19
```

Below is the regression plot for the model which had predictor “chas”

It can be seen from the residual vs fitted plot that there is a non-linear relationship between “chas” and response variable “crim”. The Normal Q-Q plot shows that the residuals are not normally distributed. This makes sense as “chas” is a qualitative variable.



(b) Fit a multiple regression model to predict the response using all of the predictors. Describe your results. For which predictors can we reject the null hypothesis Ho: βj = 0?

Assuming alpha to be 5%, we can reject the null hypothesis for the predictors whose p-values is less than 5%. Hence, we can reject null hypothesis for: “zn”, “dis”, ”rad”, ”black” and “medv”.

```
> summary(lm_multi)

Call:
lm(formula = crim ~ ., data = Boston)

Residuals:
   Min     1Q Median     3Q    Max 
-9.924 -2.120 -0.353  1.019 75.051 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  17.033228   7.234903   2.354 0.018949 *  
zn            0.044855   0.018734   2.394 0.017025 *  
indus        -0.063855   0.083407  -0.766 0.444294    
chas         -0.749134   1.180147  -0.635 0.525867    
nox         -10.313535   5.275536  -1.955 0.051152 .  
rm            0.430131   0.612830   0.702 0.483089    
age           0.001452   0.017925   0.081 0.935488    
dis          -0.987176   0.281817  -3.503 0.000502 ***
rad           0.588209   0.088049   6.680 6.46e-11 ***
tax          -0.003780   0.005156  -0.733 0.463793    
ptratio      -0.271081   0.186450  -1.454 0.146611    
black        -0.007538   0.003673  -2.052 0.040702 *  
lstat         0.126211   0.075725   1.667 0.096208 .  
medv         -0.198887   0.060516  -3.287 0.001087 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.439 on 492 degrees of freedom
Multiple R-squared:  0.454,	Adjusted R-squared:  0.4396 
F-statistic: 31.47 on 13 and 492 DF,  p-value: < 2.2e-16
```

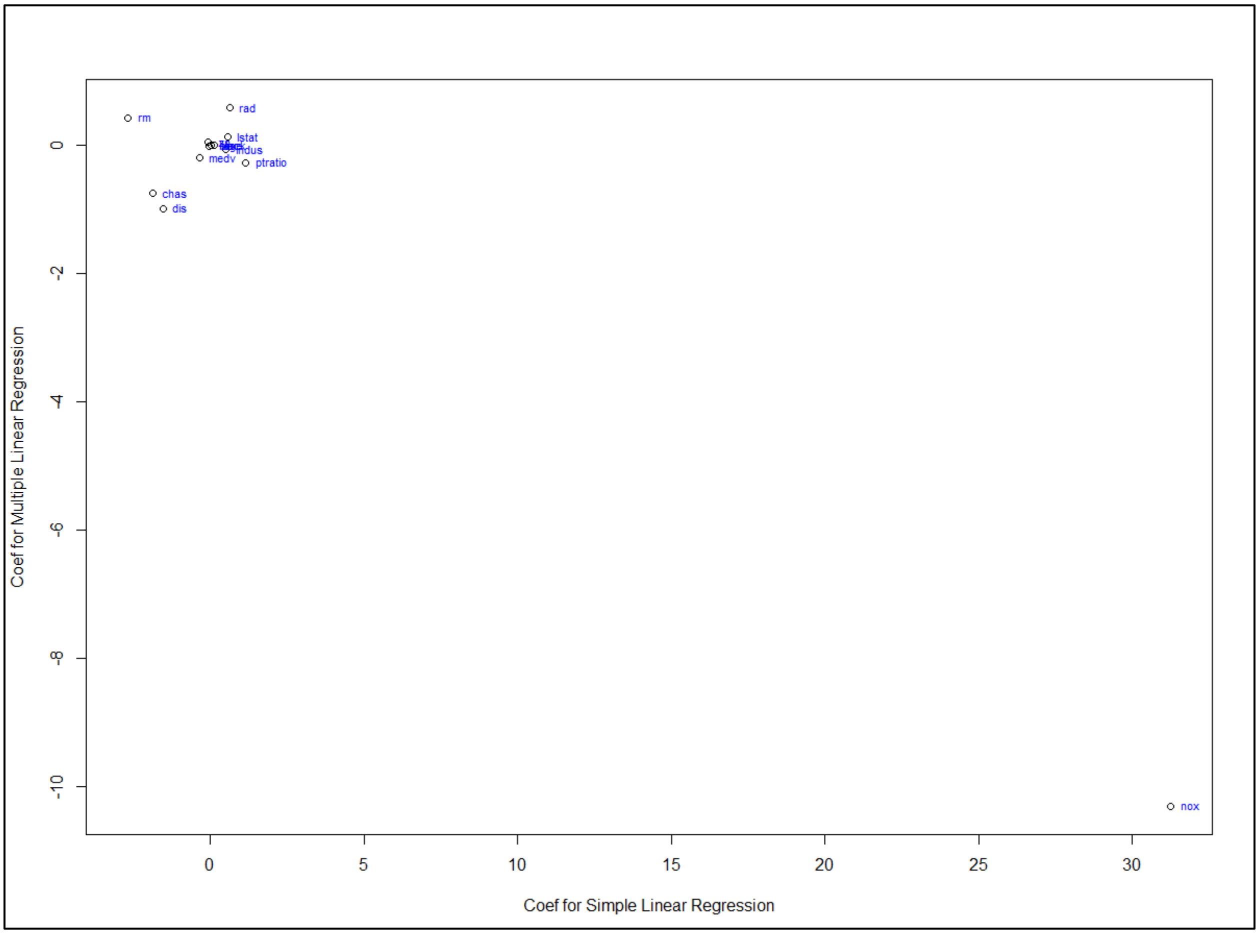
(c) How do your results from (a) compare to your results from (b)? Create a plot displaying the univariate regression coefficients from (a) on the x-axis, and the multiple regression coefficients from (b) on the y-axis. That is, each predictor is displayed as a single point in the plot. Its coefficient in a simple linear regression model is shown on the x-axis, and its coefficient estimate in the multiple linear regression model is shown on the y-axis.

Observations after comparing the results from (a) and (b) are as follows;

- Predictor “chas” is not statistically significant in both results
- Predictors “indus”, ”nox”, ”rm”, “age”, “tax”, “ptratio” and “lstat” are statistically significant for simple linear regression but not for multiple linear regression

This difference is because in the multiple regression the coefficients are correlated and in simple linear regression the response is only dependent on single independent predictor.

Below plot shows each predictor as a single plot;



(d) Is there evidence of non-linear association between any of the predictors and the response? To answer this question, for each predictor X, fit a model of the form Y = β0 + β1X + β2X2 + β3X3 + Ɛ.

Predictors “zn”, “rm”, “rad”, “tax”, “black”, “lstat” have p-values which are not statistically significant but Variables “dis”, ”rad”, ”black” and “medv” have quandratic and cubic coefficients p-values less than 5% (assuming alpha = 5%) and hence are statistically significant which means non-linear effect is visible.

This is also shown in the below summary of all the variables:

(Note: Each predictor is shown 3 times as it represents 3-degree polynomial. For e.g. first 3 rows are “zn” as 1st row represents polynomial with degree 1, 2nd row of “zn” represents polynomial with degree 2 and 3nd row of “zn” represents polynomial with degree 3)

```
> print(sum_nl)
   Predictor          Estimate        Std_error           t_value              p_value
1         zn -38.7498352143429 8.37220717285403  -4.62838943355165 4.69780623880854e-06
2         zn  23.9398319819671 8.37220717285411   2.85944094403076  0.00442050690870274
3         zn -10.0718681276877 8.37220717285406  -1.20301229051577    0.229538620491058
4      indus  78.5908191761241 7.42312095470352   10.5873014404173 8.85424265543496e-24
5      indus -24.3947964321387 7.42312095470354  -3.28632613977296  0.00108605713680062
6      indus -54.1297629340319 7.42312095470356  -7.29204916157715 1.19640469153317e-12
7        nox  81.3720154835587 7.23360503019171   11.2491648554112 2.45749078247418e-26
8        nox -28.8285942921083  7.2336050301917  -3.98537025062652 7.73675464939025e-05
9        nox -60.3618943369193 7.23360503019172  -8.34464891087915  6.9611100342705e-16
10        rm -42.3794416993869 8.32967578502905  -5.08776605393881 5.12804838748863e-07
11        rm   26.576769998347 8.32967578502904   3.19061277824445  0.00150854548563956
12        rm -5.51034200537751 8.32967578502904 -0.661531390607215    0.508575109404836
13       age  68.1820087886353 7.83970265051263   8.69701464814829 4.87880300225115e-17
14       age  37.4844703846942 7.83970265051263   4.78136379091408 2.29115552484841e-06
15       age  21.3532069846943 7.83970265051262   2.72372664329279  0.00667991535096612
16       dis -73.3885896794169   7.331478998213  -10.0100661404479 1.25324918497518e-21
17       dis  56.3730356004305   7.331478998213   7.68917644232099 7.86976668301466e-14
18       dis -42.6218774031266 7.33147899821301  -5.81354422668541 1.08883202821445e-08
19       rad  120.907445768733 6.68240174645161   18.0934116738694 1.05321131813638e-56
20       rad  17.4922987603441  6.6824017464516   2.61766643551962  0.00912055797292577
21       rad  4.69845672514562 6.68240174645159   0.70310898736978    0.482313774035658
22       tax  112.645827115381  6.8537073690599   16.4357509081734 6.97631356496829e-49
23       tax  32.0872509619434 6.85370736905989   4.68173635582938 3.66534762329281e-06
24       tax -7.99681123775891 6.85370736905992  -1.16678620885674    0.243850681055567
25   ptratio  56.0452294727408 8.12158302699871    6.900776521822 1.56548404181813e-11
26   ptratio  24.7748242612162 8.12158302699867   3.05049202585962  0.00240546785935064
27   ptratio  -22.279736819769 8.12158302699873  -2.74327514053653  0.00630051363404586
28     black -74.4311985868883 7.95464273514312  -9.35695053381289 2.73008174791173e-19
29     black  5.92641883518133 7.95464273514312  0.745026399362824    0.456604413926252
30     black -4.83456546848473 7.95464273514313 -0.607766511891969    0.543617181726899
31     lstat  88.0696661484618 7.62943609302887   11.543404397729 1.67807172578562e-27
32     lstat  15.8881643007844 7.62943609302888   2.08248212673302   0.0378041809094279
33     lstat  -11.574022255841 7.62943609302888  -1.51702198101067    0.129890587251 97
34      medv -75.0576054570241  6.5691520012923  -11.4257678072084 4.93081829258389e-27
35      medv  88.0862105806024  6.5691520012923   13.4090687143902 2.92857691192942e-35
36      medv -48.0334345541105  6.5691520012923  -7.31196881190468 1.04651002433606e-12
```

3) Suppose we collect data for a group of students in a statistics class with variables X1 =hours studied, X2 = undergrad GPA, and Y = receive an A. We fit a logistic regression and produce estimated coefficient, β0 = −6, β1 = 0.05, β2 = 1.

Ans:

We know that Logistic regression is given as

Log(p/(1-p)) = $\beta_0 + \beta_1 x_1 + \beta_2 x_2$

$Given: \beta_0 = -6 , \beta_1 = 0.05 \; and \; \beta_2 = 1$

Therefore,

Log(p/(1-p)) = $-6 + 0.05x_1 + x_2$

(a) Estimate the probability that a student who studies for 40 h and has an undergrad GPA of 3.5 gets an A in the class.

Ans:

Given Study Hours, $x_1 = 40$ *and* undergrad GPA, $x_2 = 3.5$

Probability that class gets A is;

$$P = \frac{e^{-6+0.05x1+x2}}{1+e^{-6+0.05x1+x2}}$$

$$P = \frac{e^{-6+0.05*40+3.5}}{1+e^{-6+0.05*40+3.5}}$$

<u>P = 0.3775</u>

(b) How many hours would the student in part (a) need to study to have a 50% chance of getting an A in the class?

Ans: Given Probability of the class getting A is P=0.5 and undergrad GPA , x2= 3.5

$$\frac{e^{-6+0.05x1+3.5}}{1+e^{-6+0.05x1+3.5}} = 0.5$$

<u>X1 = 50</u>

(4) 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.

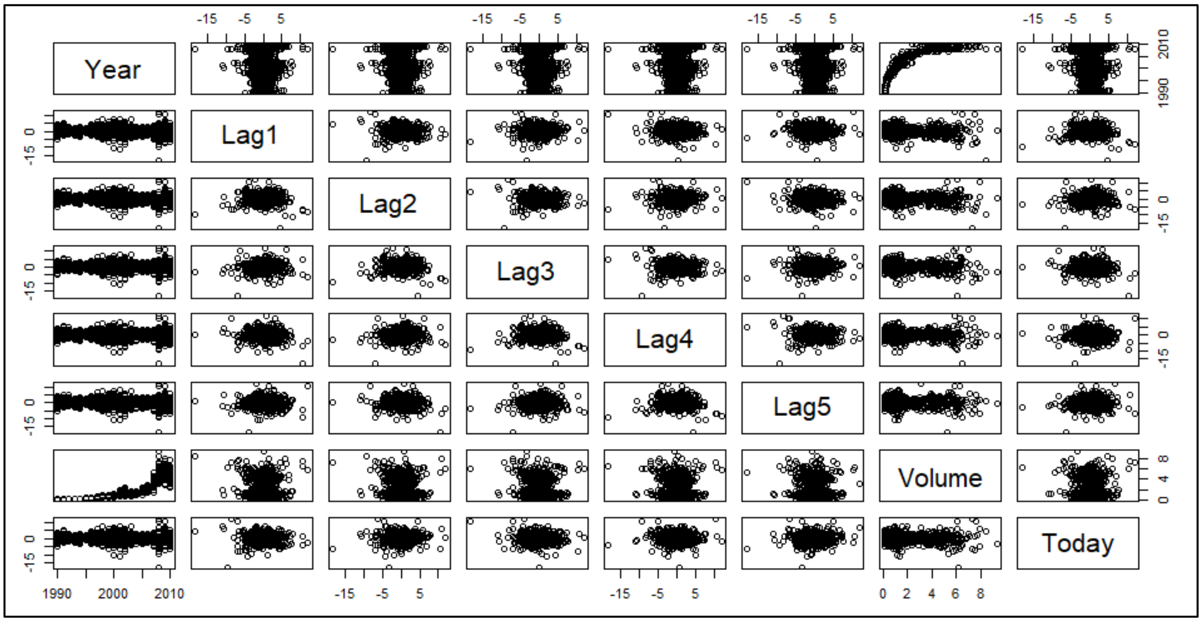
(a) Produce some numerical and graphical summaries of the Weekly data. Do there appear to be any patterns?

The pairwise correlations between the numeric variables is shown below;

```
> cor(Weekly[,-9])
              Year         Lag1        Lag2        Lag3         Lag4         Lag5      Volume        Today
Year    1.00000000 -0.032289274 -0.03339001 -0.03000649 -0.031127923 -0.030519101  0.84194162 -0.032459894
Lag1   -0.03228927  1.000000000 -0.07485305  0.05863568 -0.071273876 -0.008183096 -0.06495131 -0.075031842
Lag2   -0.03339001 -0.074853051  1.00000000 -0.07572091  0.058381535 -0.072499482 -0.08551314  0.059166717
Lag3   -0.03000649  0.058635682 -0.07572091  1.00000000 -0.075395865  0.060657175 -0.06928771 -0.071243639
Lag4   -0.03112792 -0.071273876  0.05838153 -0.07539587  1.000000000 -0.075675027 -0.06107462 -0.007825873
Lag5   -0.03051910 -0.008183096 -0.07249948  0.06065717 -0.075675027  1.000000000 -0.05851741  0.011012698
Volume  0.84194162 -0.064951313 -0.08551314 -0.06928771 -0.061074617 -0.058517414  1.00000000 -0.033077783
Today  -0.03245989 -0.075031842  0.05916672 -0.07124364 -0.007825873  0.011012698 -0.03307778  1.000000000
```

- There is very little correlation between the lag variables i.e previous day's returns and today's returns as their values are close to zero.
- The only correlation that we see is between Year and Volume.

Below plot of correlation also shows that volume increases over time;



(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 of the logistic function is given below;

```
> summary(glm.fits)

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
```

Only Predictor “Lag2” appears to be statistically significant.

(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.

Confusion matrix for above logistic regression is given below;

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

glm.pred Down  Up
    Down   54  48
    Up    430 557
> mean(glm.pred==Weekly$Direction)
[1] 0.5610652
>
```

The confusion matrix can tell us how many values are correctly and incorrectly predicted. The diagonal elements in the matrix are the correct predictions i.e. For 54 days the Directions are predicted as "Down" are actually "Down" and for 557 days of Directions which were predicted as "Up" are actually "Up". The off-diagonal represents the incorrect predictions. We can calculate overall Accuracy and the error rate of the model using confusion matrix.

The overall fraction of correctly predicted = 54+557/1089 = 56.106%

(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).

Confusion matrix and the overall fraction of correct predictions for the data from 2009 to 2010 is given below;

```
> table(glm_tst,Week_tst$Direction)

glm_tst Down Up
   Down    9  5
   Up     34 56
> mean(glm_tst==Week_tst$Direction)
[1] 0.625
```

Overall Fraction of correct prediction = 62.5%

(e) Repeat (d) using LDA.

Confusion matrix and the overall fraction of correct predictions (i.e. 62.5%)using LDA for the data from 2009 to 2010 is given below;

```
> table(lda_tst$class,Week_tst$Direction)

       Down Up
  Down    9  5
  Up     34 56
> mean(lda_tst$class==Week_tst$Direction)
[1] 0.625
```

(f) Repeat (d) using QDA.

Confusion matrix and the overall fraction of correct predictions using QDA for the data from 2009 to 2010 is given below;

```
> table(qda_tst$class,week_tst$Direction)

       Down Up
  Down    0  0
  Up     43 61
> mean(qda_tst$class==week_tst$Direction)
[1] 0.5865385
```

Overall Fraction of correct prediction = 58.65%
Even though the error rate is 41.35%, the model is not performing well as it is classifying all of the data to be "Up".

(g) Repeat (d) using KNN with K = 1.

Confusion matrix and the overall fraction of correct predictions using K-n-n for the data from 2009 to 2010 is given below;

```
> table(knn_trn,week_tst$Direction)

knn_trn Down Up
   Down   21 29
   Up     22 32
> mean(knn_trn==week_tst$Direction)
[1] 0.5096154
```

Overall Fraction of correct prediction = 50.9%
Even though the overall Accuracy is less than other models, unlike other models it can identify approximately 49% of "Down" directions i.e. True negatives correctly

(h) Which of these methods appears to provide the best results on this data?

By comparing the above model LDA and Logistic Regression would give better results as have performed well on the test data compared to others.

(i) Experiment with different combinations of predictors, including possible transformations and interactions, for each of the methods. Report the variables, method, and associated confusion matrix that appears to provide the best results on the held out data. Note that you should also experiment with values for K in the KNN classifier.

Ans:

Different combinations of predictors and methods were tried and below are the results of their confusion matrix and overall fraction of correct predictions;

1) Logistic regression

(1) Variables: Lag1 and (Lag2)^2

```
> table(glm_tst_1,week_tst$Direction)

glm_tst_1 Down Up
     Down    6  6
     Up     37 55
> mean(glm_tst_1==week_tst$Direction)
[1] 0.5865385
```

(2) Variables: Lag2 and Lag 5

```
> table(glm_tst_2,week_tst$Direction)

glm_tst_2 Down Up
     Down    7  5
     Up     36 56
> mean(glm_tst_2==week_tst$Direction)
[1] 0.6057692
```

(3) Variables: Lag1, Lag2, and interaction between Lag3 and Lag4(Lag3:Lag4)

```
> table(glm_tst_3,week_tst$Direction)

glm_tst_3 Down Up
     Down    8  7
     Up     35 54
> mean(glm_tst_3==week_tst$Direction)
[1] 0.5961538
```

2) LDA

(1) Variables: Lag2 and interaction term between Lad2 and Lag 5 (Lag2*Lag5)

```
> table(lda_tst_1$class,week_tst$Direction)

        Down Up
  Down     6  5
  Up      37 56
> mean(lda_tst_1$class==week_tst$Direction)
[1] 0.5961538
```

(2) Variables: Lag1, Lag2, Lag3 and Volume

```
> table(lda_tst_2$class,week_tst$Direction)

        Down Up
  Down    30 37
  Up      13 24
> mean(lda_tst_2$class==week_tst$Direction)
[1] 0.5192308
```

(3) Variables: Lag2 and (Volume)^3

```
> table(lda_tst_3$class,Week_tst$Direction)

       Down Up
  Down   14 14
  Up     29 47
> mean(lda_tst_3$class==Week_tst$Direction)
[1] 0.5865385
```

3) QDA

(1) Variables: Lag1, Lag2 and poly(Volume,3)

```
> table(qda_tst_1$class,Week_tst$Direction)

       Down Up
  Down   39 53
  Up      4  8
> mean(qda_tst_1$class==Week_tst$Direction)
[1] 0.4519231
```

(2) Variables: Lag1, (Lag2)^2, Lag3, interaction between Lag4 and Lag5 (Lag4:Lag5)

```
> table(qda_tst_2$class,Week_tst$Direction)

       Down Up
  Down    5 13
  Up     38 48
> mean(qda_tst_2$class==Week_tst$Direction)
[1] 0.5096154
```

(3) Variables: Lag2 and Lag5

```
> table(qda_tst_3$class,Week_tst$Direction)

       Down Up
  Down    3 11
  Up     40 50
> mean(qda_tst_3$class==Week_tst$Direction)
[1] 0.5096154
```

4) KNN

(1) Variables: Lag2, Lag3 and Lag4; K= 3

```
> table(knn_trn1,Week_tst$Direction)

knn_trn1 Down Up
    Down   17 26
    Up     26 35
> mean(knn_trn1==Week_tst$Direction)
[1] 0.5
```

(2) Variables: Lag2, Lag4 and Volume; K= 5

```
> table(knn_trn2,Week_tst$Direction)

knn_trn2 Down Up
    Down   21 33
    Up     22 28
> mean(knn_trn2==Week_tst$Direction)
[1] 0.4711538
```

(3) Variables: Lag1 and Lag5; K= 4

```
> table(knn_trn3,Week_tst$Direction)

knn_trn3 Down Up
    Down   17 25
    Up     26 36
> mean(knn_trn3==Week_tst$Direction)
[1] 0.5096154
```