Midterm

Problem 1

Reading the data

```
library(readr)
JobProf <- read_csv("E:/SUBJECTS/569 MATH SL S17--/midterm/JobProf.txt")
```

```
## Parsed with column specification:
## cols(
##   Y = col_integer(),
##   X1 = col_integer(),
##   X2 = col_integer(),
##   X3 = col_integer(),
##   X4 = col_integer()
## )
```

a. Full model

```
all_lm=lm(Y~.,data=JobProf)
summary(all_lm)
```

```
##
## Call:
## lm(formula = Y ~ ., data = JobProf)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.9779 -3.4506  0.0941  2.4749  5.9959
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -124.38182    9.94106 -12.512 6.48e-11 ***
## X1             0.29573    0.04397   6.725 1.52e-06 ***
## X2             0.04829    0.05662   0.853  0.40383
## X3             1.30601    0.16409   7.959 1.26e-07 ***
## X4             0.51982    0.13194   3.940  0.00081 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.099 on 20 degrees of freedom
## Multiple R-squared:  0.9629, Adjusted R-squared:  0.9555
## F-statistic: 129.7 on 4 and 20 DF,  p-value: 5.262e-14
```

Estimation sigma-hat-square

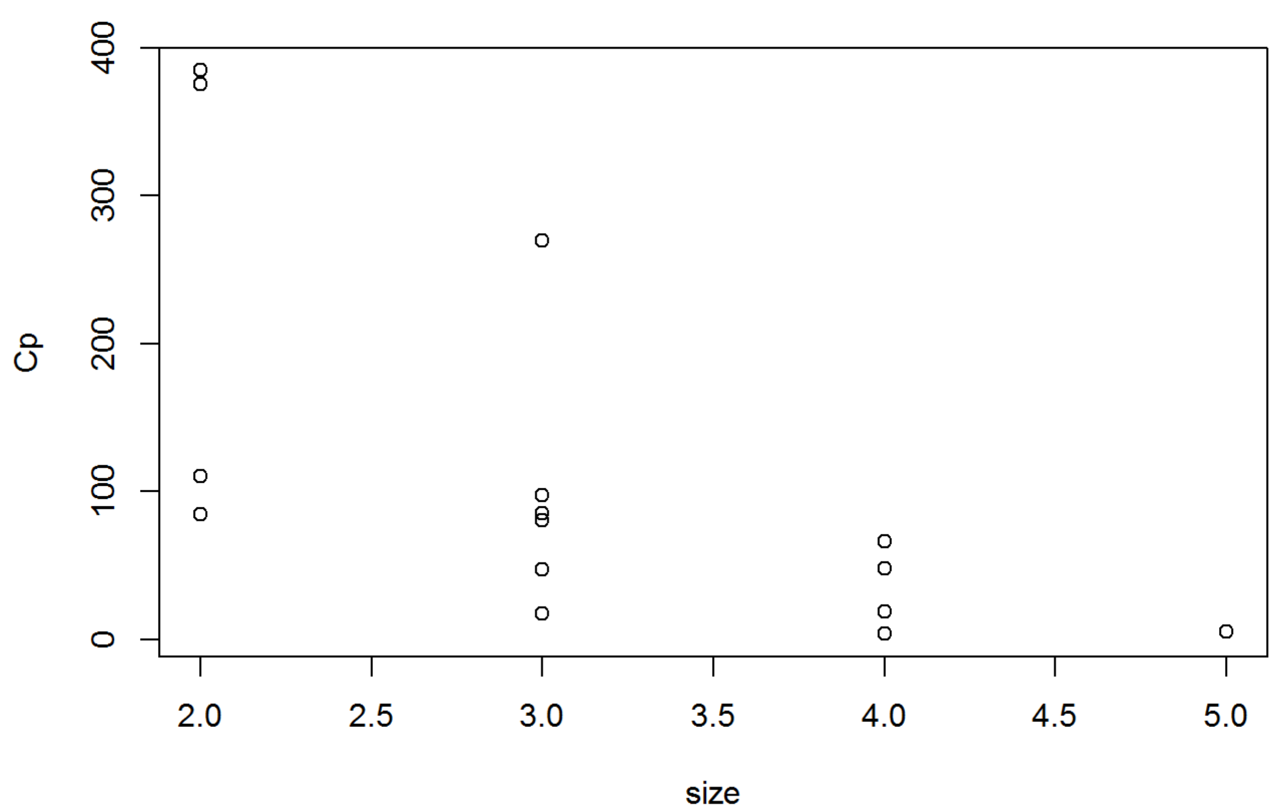
```
(sigma(all_lm))**2
```

```
## [1] 16.79888
```

Except X2 all the other predictor variables are statistically significant at the level α = 0.05.

(b)Best subset

```
library(leaps)
best_subs=leaps(x=as.matrix(JobProf[,-1]),y=as.matrix(JobProf[,1]))
plot(x=best_subs$size,y=best_subs$Cp,xlab='size',ylab='Cp')
```




```
min(best_subs$Cp)
```

```
## [1] 3.727399
```

```
best_subs$which[which(best_subs$Cp==(min(best_subs$Cp))),]
```

```
##     1     2     3     4
##  TRUE FALSE  TRUE  TRUE
```

```
fit_bestsub<-lm(Y~X1+X3+X4,data = JobProf)
summary(fit_bestsub)
```

```
##
## Call:
## lm(formula = Y ~ X1 + X3 + X4, data = JobProf)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.4579 -3.1563 -0.2057  1.8070  6.6083
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -124.20002    9.87406 -12.578 3.04e-11 ***
## X1            0.29633    0.04368   6.784 1.04e-06 ***
## X3            1.35697    0.15183   8.937 1.33e-08 ***
## X4            0.51742    0.13105   3.948 0.000735 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.072 on 21 degrees of freedom
## Multiple R-squared:  0.9615, Adjusted R-squared:  0.956
## F-statistic:   175 on 3 and 21 DF,  p-value: 5.16e-15
```

The model of form Y~X1+X3+X4 has the minimum Cp at 3.727399

c. Forward and Backward selection

```
fit0<-lm(Y~1,data=JobProf)
fit.forward<-step(fit0,scope=list(lower=Y~1, upper=Y~X1+X2+X3+X4),direction='forward')
```

```
## Start:  AIC=149.3
## Y ~ 1
##
##        Df Sum of Sq    RSS    AIC
## + X3    1    7286.0 1768.0 110.47
## + X4    1    6843.3 2210.7 116.06
## + X1    1    2395.9 6658.1 143.62
## + X2    1    2236.5 6817.5 144.21
## <none>              9054.0 149.30
##
## Step:  AIC=110.47
## Y ~ X3
##
##        Df Sum of Sq     RSS     AIC
## + X1    1   1161.37  606.66  85.727
## + X4    1    656.71 1111.31 100.861
## <none>              1768.02 110.469
## + X2    1     12.21 1755.81 112.295
##
## Step:  AIC=85.73
## Y ~ X3 + X1
##
##        Df Sum of Sq    RSS    AIC
## + X4    1   258.460 348.20 73.847
## <none>              606.66 85.727
## + X2    1     9.937 596.72 87.314
##
## Step:  AIC=73.85
## Y ~ X3 + X1 + X4
##
##        Df Sum of Sq    RSS    AIC
## <none>              348.20 73.847
## + X2    1     12.22 335.98 74.954
```

```
summary(fit.forward)
```

```
##
## Call:
## lm(formula = Y ~ X3 + X1 + X4, data = JobProf)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.4579 -3.1563 -0.2057  1.8070  6.6083
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -124.20002    9.87406 -12.578 3.04e-11 ***
## X3             1.35697    0.15183   8.937 1.33e-08 ***
## X1             0.29633    0.04368   6.784 1.04e-06 ***
## X4             0.51742    0.13105   3.948 0.000735 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.072 on 21 degrees of freedom
## Multiple R-squared:  0.9615, Adjusted R-squared:  0.956
## F-statistic:   175 on 3 and 21 DF,  p-value: 5.16e-15
```

```
fit.backward<-step(all_lm,scope=list(lower=Y~1, upper=Y~X1+X2+X3+X4),direction='backward')
```

```
## Start:  AIC=74.95
## Y ~ X1 + X2 + X3 + X4
##
##        Df Sum of Sq     RSS     AIC
## - X2    1     12.22  348.20  73.847
## <none>               335.98  74.954
## - X4    1    260.74  596.72  87.314
## - X1    1    759.83 1095.81 102.509
## - X3    1   1064.15 1400.13 108.636
##
## Step:  AIC=73.85
## Y ~ X1 + X3 + X4
##
##        Df Sum of Sq     RSS     AIC
## <none>               348.20  73.847
## - X4    1    258.46  606.66  85.727
## - X1    1    763.12 1111.31 100.861
## - X3    1   1324.39 1672.59 111.081
```

```
summary(fit.backward)
```

```
##
## Call:
## lm(formula = Y ~ X1 + X3 + X4, data = JobProf)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.4579 -3.1563 -0.2057  1.8070  6.6083
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -124.20002    9.87406 -12.578 3.04e-11 ***
## X1             0.29633    0.04368   6.784 1.04e-06 ***
## X3             1.35697    0.15183   8.937 1.33e-08 ***
## X4             0.51742    0.13105   3.948 0.000735 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.072 on 21 degrees of freedom
## Multiple R-squared:  0.9615, Adjusted R-squared:  0.956
## F-statistic:   175 on 3 and 21 DF,  p-value: 5.16e-15
```

Same optimal model is returned, Y ~ X1 + X3 + X4

```
anova(fit.forward,fit.backward)
```

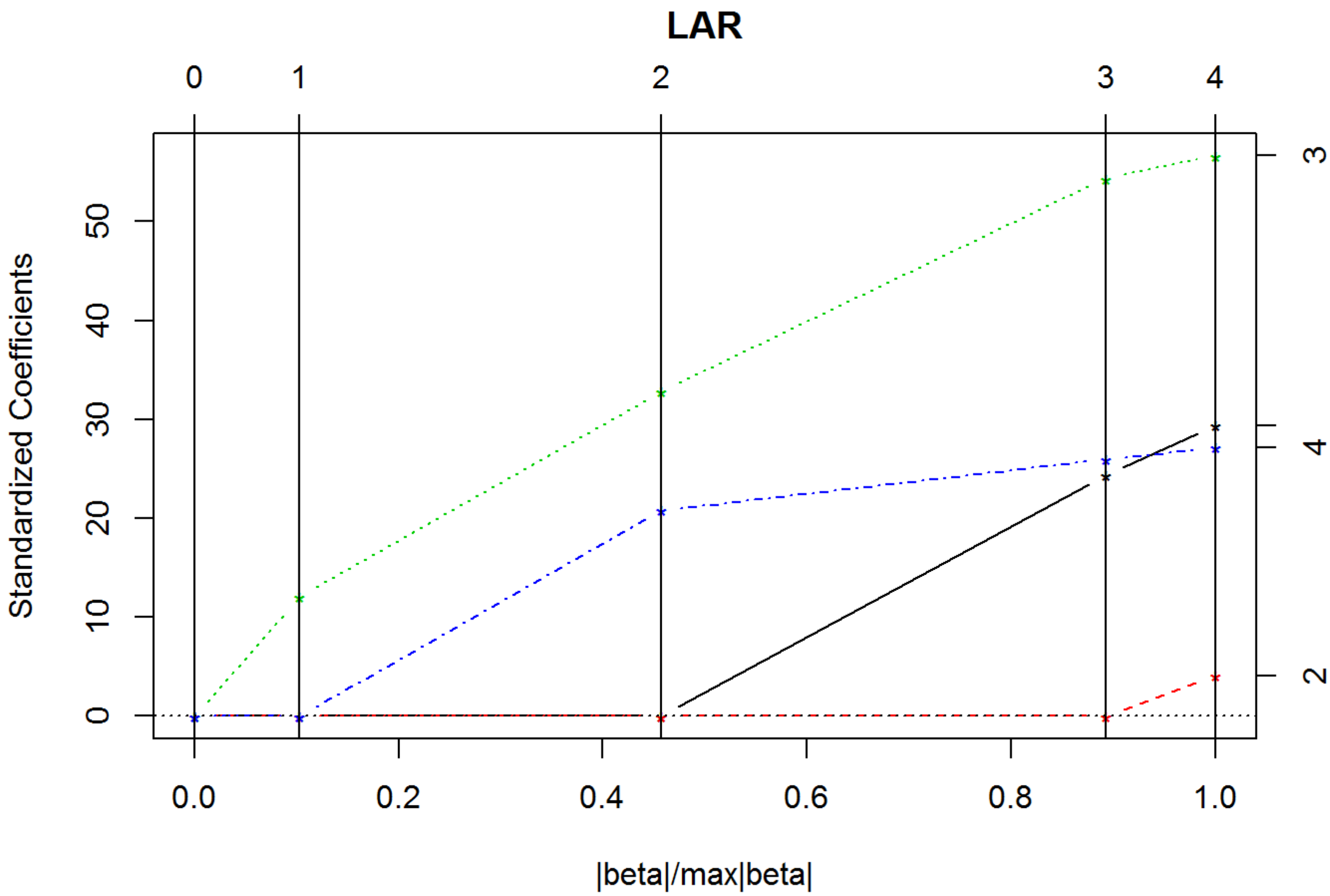
```
## Analysis of Variance Table
##
## Model 1: Y ~ X3 + X1 + X4
## Model 2: Y ~ X1 + X3 + X4
##   Res.Df   RSS Df Sum of Sq F Pr(>F)
## 1     21 348.2
## 2     21 348.2  0         0
```

d. Lasso and LAR

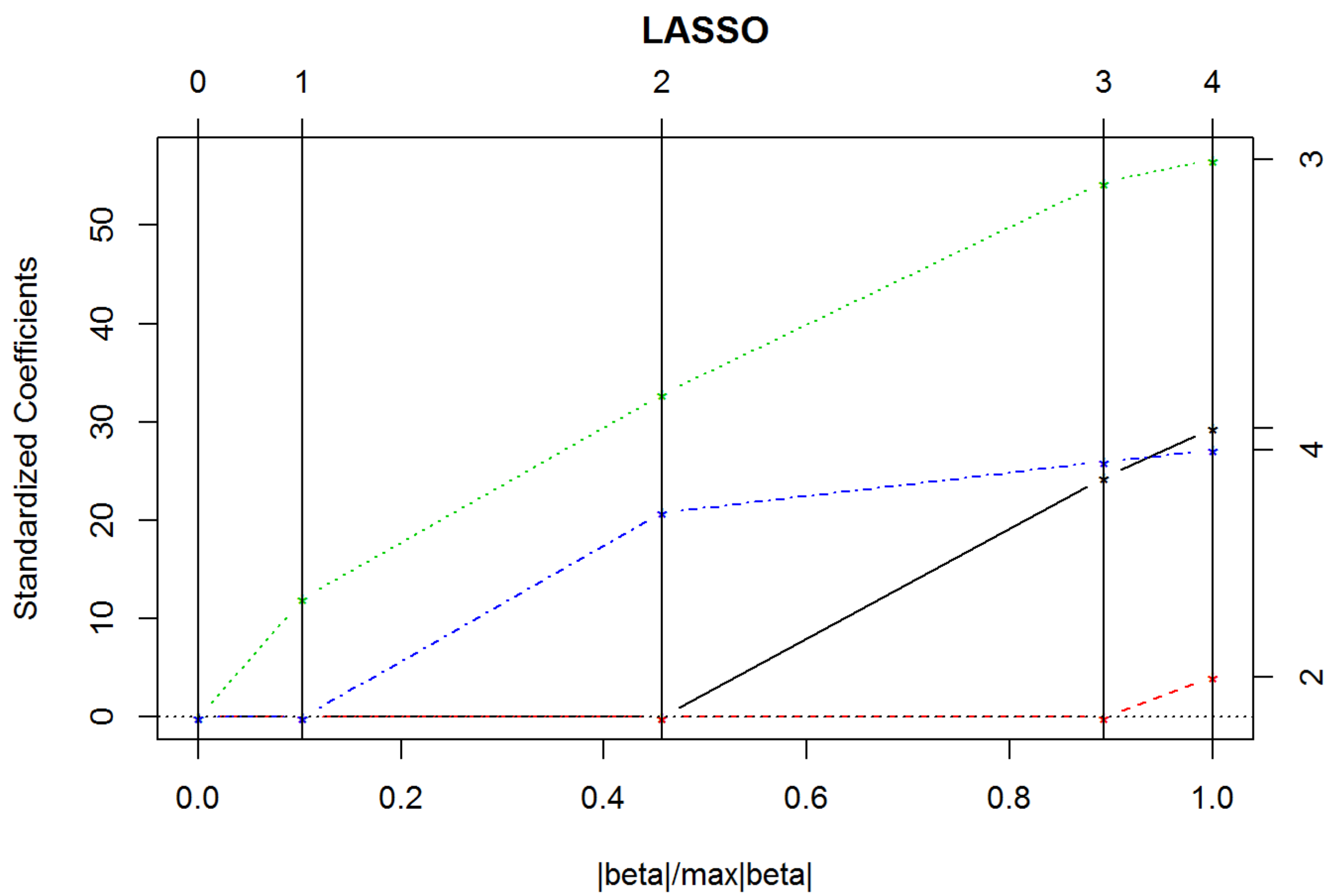
```
library(lars)
```

```
## Loaded lars 1.2
```

```
fit.lars<-lars(x=as.matrix(JobProf[,-1]),y=as.matrix(JobProf[,1]), type="lar")
plot(fit.lars)
```




```
fit.lasso<-lars(x=as.matrix(JobProf[,-1]),y=as.matrix(JobProf[,1]), type="lasso")
plot(fit.lasso)
```



There is no difference between the two profiles

Problem 2

Reading the data

```
library(readr)
Car <- read_csv("E:/SUBJECTS/569 MATH SL S17--/midterm/Car.txt")
```

```
## Parsed with column specification:
## cols(
##   Y = col_integer(),
##   X1 = col_integer(),
##   X2 = col_integer()
## )
```

a. Maximum likelihood estimates

```
fit<-glm(Y~.,family=binomial(link='logit'),data=Car)
summary(fit)
```

```
##
## Call:
## glm(formula = Y ~ ., family = binomial(link = "logit"), data = Car)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6189  -0.8949  -0.5880   0.9653   2.0846
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -4.73931    2.10195  -2.255   0.0242 *
## X1           0.06773    0.02806   2.414   0.0158 *
## X2           0.59863    0.39007   1.535   0.1249
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 44.987  on 32  degrees of freedom
## Residual deviance: 36.690  on 30  degrees of freedom
## AIC: 42.69
##
## Number of Fisher Scoring iterations: 4
```

b. Prediction

```
pre=data.frame(X1=50,X2=3)
predict(fit,newdata = pre)
```

```
##         1
## 0.4432137
```

Problem 3

Reading data

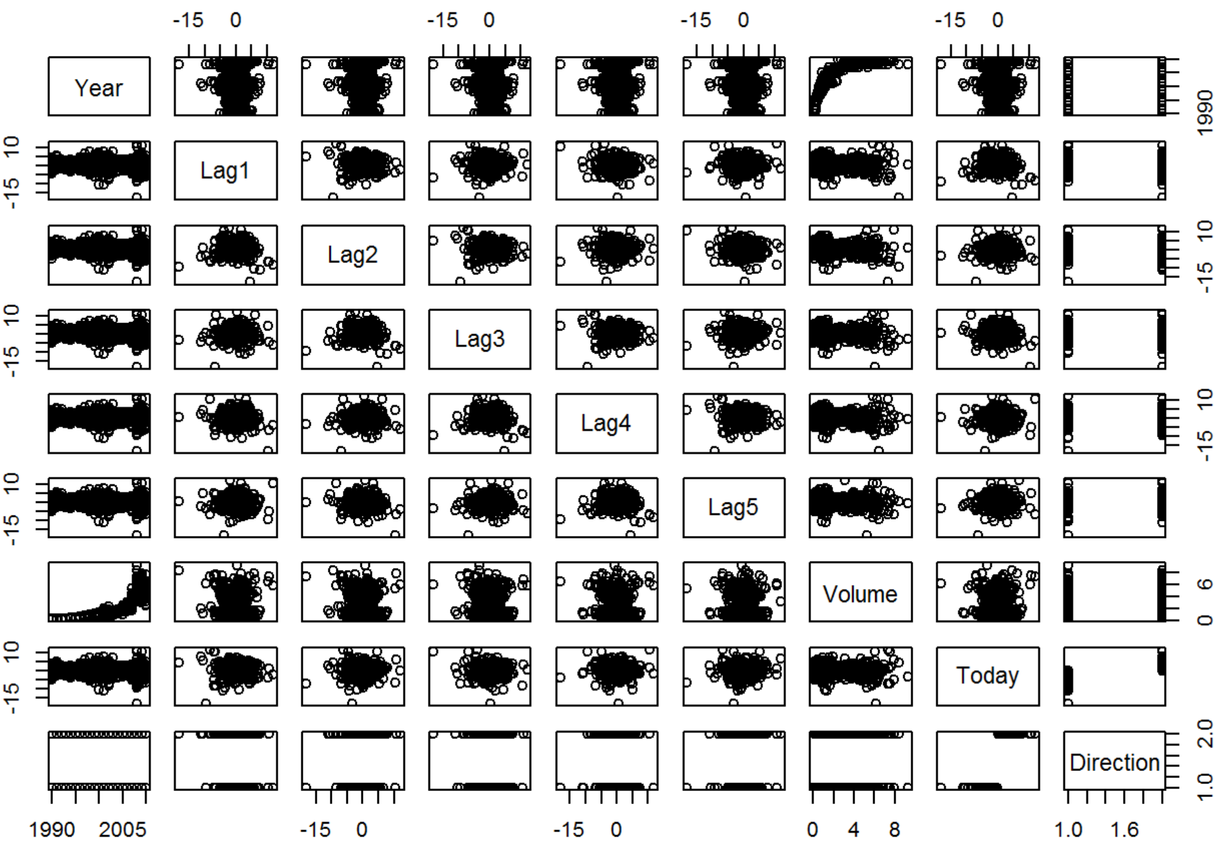
```
library(ISLR)
```

```
## Warning: package 'ISLR' was built under R version 3.3.3
```

```
data("Weekly")
```

a. Patterns in the data

```
pairs(Weekly)
```



There seems to be a positive linear relationship between Volume and Year

(b)Logistic Regression

```
fit_weekly_logistic=glm(Direction~Lag1+Lag2+Lag3+Lag4+Lag5+Volume,family = binomial(link = 'logi
t'),data=Weekly)
summary(fit_weekly_logistic)
```

```
##
## Call:
## glm(formula = Direction ~ Lag1 + Lag2 + Lag3 + Lag4 + Lag5 +
##     Volume, family = binomial(link = "logit"), 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
```

The Intercept and Lag2 are statistically signnificant at .001 and .01 levels respectively

c. Logistic Regression with 1990 to 2008 data

Separating the data

```
condition<-Weekly$Year<2009
train<-Weekly[condition,]
test<-Weekly[!condition,]
stopifnot(nrow(train)+nrow(test)==nrow(Weekly))
```

Fit the model

```
fit_weekly_logistic_lag2=glm(Direction~Lag2,family = binomial(link = 'logit'),data = train)
summary(fit_weekly_logistic_lag2)
```

```
##
## Call:
## glm(formula = Direction ~ Lag2, family = binomial(link = "logit"),
##     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
```

Prediction

```
prob<-predict(fit_weekly_logistic_lag2,newdata = test,type = 'response')
predictions<-rep("Up",nrow(test))
predictions[prob<0.5]="Down"
```

Confusion Matrix

```
predictions<-as.matrix(predictions)
target<-test['Direction']
target<-as.matrix(target)
table(predictions,target)
```

```
##            target
## predictions Down Up
##        Down    9  5
##        Up     34 56
```

Correct predictions

```
mean(predictions==target)
```

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

d. LDA

Fit the model

```r
library(MASS)
fit_lda=lda(Direction~Lag2,data=train)
```

Predictions

```r
pred_lda<-predict(fit_lda,newdata = test)
```

Confusion Matrix

```r
table(pred_lda$class,target)
```

```
##       target
##        Down Up
##   Down    9  5
##   Up     34 56
```

Correct predictions

```r
mean(pred_lda$class==target)
```

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

e. QDA

Fit the model

```r
fit_qda<-qda(Direction~Lag2,data=train)
```

Predictions

```r
pred_qda<-predict(fit_qda,newdata=test)
```

Confusion Matrix

```r
table(pred_qda$class,target)
```

```
##       target
##        Down Up
##   Down    0  0
##   Up     43 61
```

Correct Predictions

```r
mean(pred_qda$class==target)
```

```
## [1] 0.5865385
```

(f)KNN

Train and test sets

```
train1<-as.matrix(train['Lag2'])
train_target<-as.matrix(train['Direction'])
stopifnot(nrow(train1)==nrow(train_target))
test1<-as.matrix(test['Lag2'])
stopifnot(nrow(test1)==length(target))
```

Fit the model and predict

```
library(class)
set.seed(3985)
fit_knn<-knn(train=train1,test=test1,cl=train_target,k=1)
```

Confusion Matrix

```
table(fit_knn,target)
```

```
##        target
## fit_knn Down Up
##    Down   21 30
##    Up     22 31
```

Correct Predictions

```
mean(fit_knn==target)
```

```
## [1] 0.5
```

g. Best results

We can get the best results in LDA and logistic regression with correct predictions for 62.5%