# Homework 2 - 4620

*Tyler Poelking*

*9/18/2017*

Problem 1: a) When the Bayes decision boundry is linear, LDA is expected to perform better on both the training and the test set. This is because of the bias-variance tradeoff. QDA involves higher variance and thus will overfit to the data. In this scenario, the QDA decision boundary is inferior, because it suffers from higher variance without a corresponding decrease in bias.

b) Contrastly, when the Bayes decision boundry is non-linear, QDA is expected to perform better on both the training and the test set. Again, this is because of the bias-variance tradeoff. QDA involves higher variance and can account for the varying covariance matrices between the K classes. In this scenario LDA will suffer from high bias.

c) As the sample size n increases, we expect the prediction accuracy of QDA to be superior to that of LDA. This is because, with a large training set, we do not want to have to be concerned about the variance of the classifier. This benefit outweighs the fact that QDA tends to require more time and computational power.

d) False. Unless the co-variance matrices between the training data's K classes is extremely linear, QDA will account for noise in the data and consider it reality (overfit it), thus leading to suffering performance and test error rate.

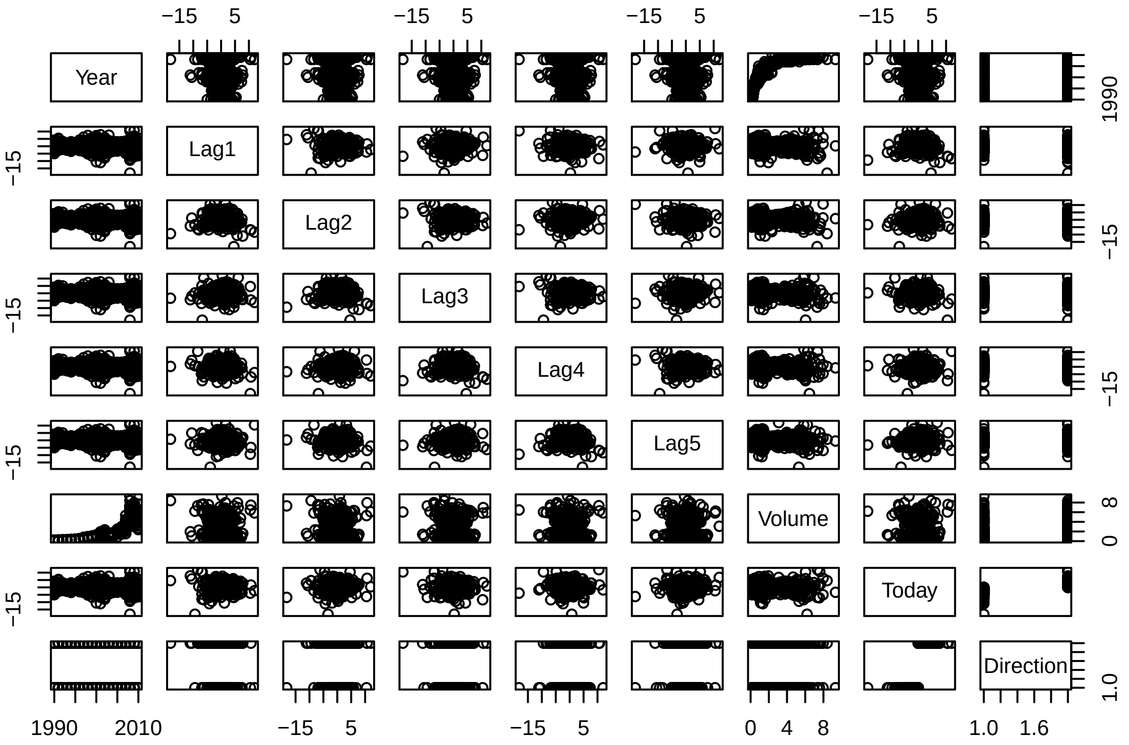
```
#Install + load package
#install.packages("ISLR")
library(ISLR)

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
##  Min.   :-18.1950   Min.   :-18.1950   Min.   :0.08747
##  1st Qu.: -1.1580   1st Qu.: -1.1660   1st Qu.:0.33202
##  Median :  0.2380   Median :  0.2340   Median :1.00268
##  Mean   :  0.1458   Mean   :  0.1399   Mean   :1.57462
##  3rd Qu.:  1.4090   3rd Qu.:  1.4050   3rd Qu.:2.05373
##  Max.   : 12.0260   Max.   : 12.0260   Max.   :9.32821
##      Today          Direction
##  Min.   :-18.1950   Down:484
##  1st Qu.: -1.1540   Up  :605
##  Median :  0.2410
##  Mean   :  0.1499
##  3rd Qu.:  1.4050
##  Max.   : 12.0260
```

```
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
```

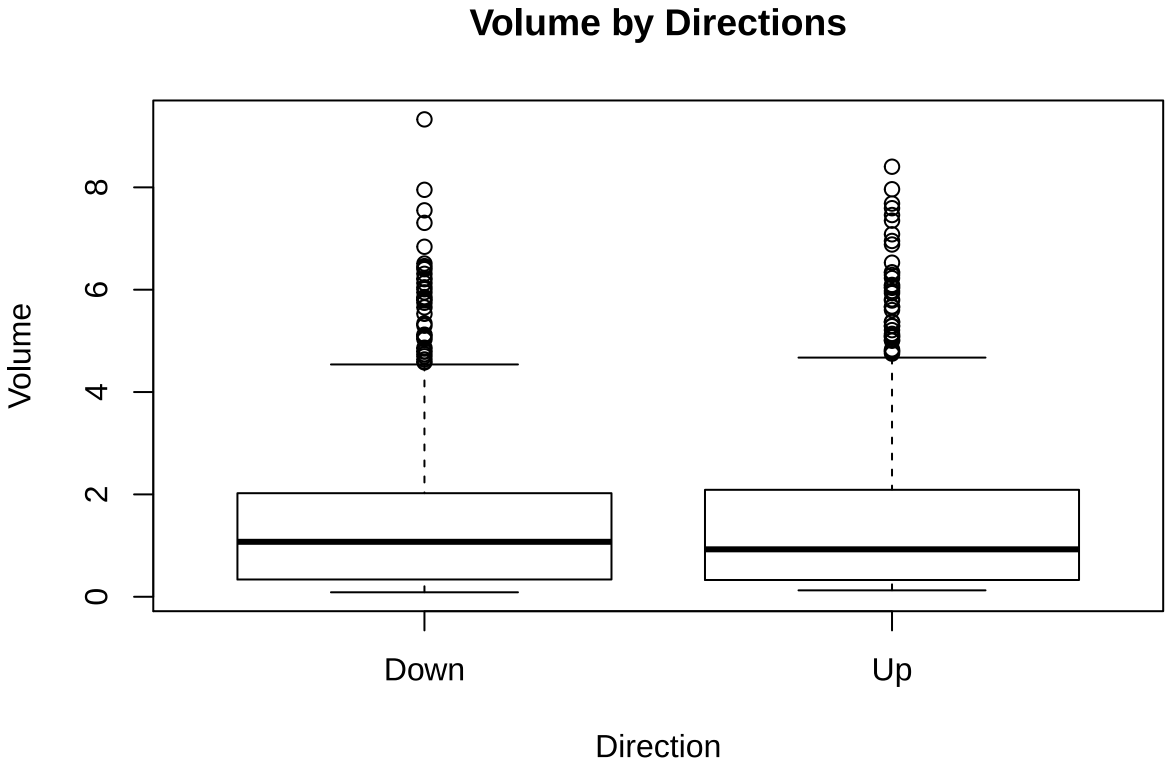
```
pairs(Weekly)
```



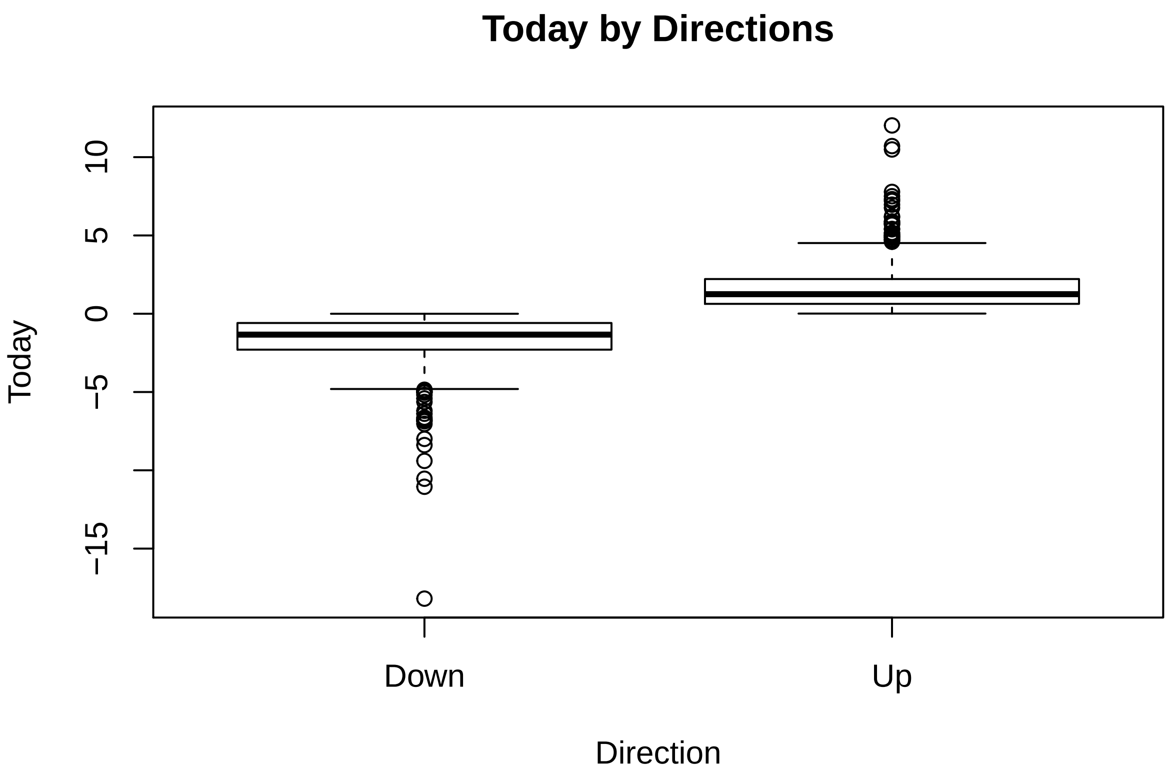
```
#remove factor vars so cor function works
cor(Weekly[, -9])
```

```
##               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
```

```
Weekly$Direction =as.factor(Weekly$Direction)
attach(Weekly)

boxplot(Volume~Direction, main="Volume by Directions",
xlab="Direction", ylab="Volume")
```



```
boxplot(Today~Direction, main="Today by Directions",
xlab="Direction", ylab="Today")
```



Year and Volume appear to be exponentially related. Specifically, as Year increases, the Volume increases at an increasing rate. The scatter plots also show that there aren't many strong correlations amongst the continuous variables in the dataset. Though it does reveal several potential outliers.

The spread of Volume between the two Directions are quite similar to each other. The spread of Today between the two Directions, however, differ. Specifically, the mean Today of the 'down' direction is smaller than the mean Today of the 'up' direction, and most of the outliers in the 'down' group are below its mean, while most the outliers in the 'up' group are above the mean.

```
##
## 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
```

One predictor, Lag2, seems to have statistical significance in predicting Direction in this model, since it has a p-value of 0.296 in the summary table above.

```
##           Direction
## model.pred Down  Up
##       Down   54  48
##       Up    430 557
```

The majority of observations belong to the 'Up' Direction class. The precision of this class is 557/(557+430) = 56.38% and the precision of the 'Down' class is 54/(54+48) = 52.9%. The recall of the 'Up' class is 557/(557+48) = 92.1% and the recall of the 'Down' class is 54/(430+54) = 11.2%. Therefore, the logistic regression is wrong the most during the weeks the market has gone down.

The percent of predictions correct is only (54+557)/(54+557+48+430) = 56.1%.

```
##      Year
## 986  2009
## 1038 2010
```

```
##     Year
## 1   1990
## 48  1991
## 100 1992
## 152 1993
## 204 1994
## 256 1995
## 308 1996
## 361 1997
## 413 1998
## 465 1999
## 517 2000
## 569 2001
## 621 2002
## 673 2003
## 725 2004
## 777 2005
## 829 2006
## 881 2007
## 934 2008
```

```
##
## model2.pred Down Up
##        Down    9  5
##        Up     34 56
```

The overall fraction of correct predictions is (9+56)/(9+5+34+56) = 65/104 = 62.5%

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

The overall fraction of correct predictions for this LDA model is (9+56)/(9+5+34+56) = 65/104 = 62.5%

Precision for 'Up': 56/(56+34) = 62.2% Precision for 'Down': 9/(9+5) = 64.3% Recall for 'Up': 56/(56+5) = 91,8% Recall for 'Down': 9/(34+9) = 20.9%

Problem 10 part h:

LDA and Logistic Regression provide the same test results so one is not better than the other under this test set.