Assignment_E

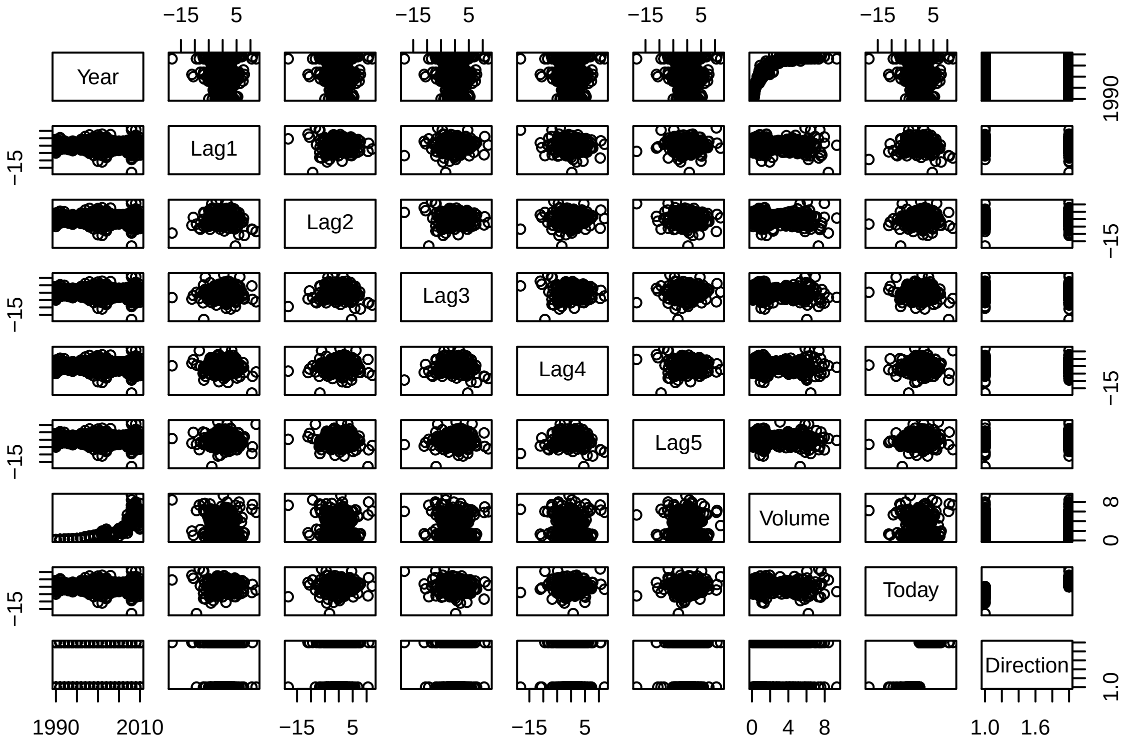
Sangamesh

3 December 2018

Q.1] Consider the Weekly data set, which is part of ISLR package. It contains the weekly stock market returns for 21 years.

a] Produce some numerical and graphical summaries of the Weekly data. Do there appear to be any pattern?

```
##       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
```



We can observe that the Weekly data from ISLR has Volume and Year taken together has logarithmic distribution.

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 appears to be statistically significant? If so, which ones?

```
##
## 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
```

Statistically significant predictor among the given is Lag2 only since the p-value is greater than the significant code attached to it.

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.

```
##       Weeklyglm.preds
##        Down  Up
##   Down   54 430
##   Up     48 557
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction Down  Up
##       Down   54 430
##       Up     48 557
##
##                Accuracy : 0.5611
##                  95% CI : (0.531, 0.5908)
##     No Information Rate : 0.9063
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.035
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.52941
##             Specificity : 0.56434
##          Pos Pred Value : 0.11157
##          Neg Pred Value : 0.92066
##              Prevalence : 0.09366
##          Detection Rate : 0.04959
##    Detection Prevalence : 0.44444
##       Balanced Accuracy : 0.54687
##
##        'Positive' Class : Down
##
```

There are a predominance of Up prediction. The model predicts well the Up direction, but it predict poorly the Down direction.

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.)

```
##       glm.preds.d
##        Down Up
##   Down    9 34
##   Up      5 56
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction Down Up
##       Down    9 34
##       Up      5 56
##
##                Accuracy : 0.625
##                  95% CI : (0.5247, 0.718)
##     No Information Rate : 0.8654
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.1414
##  Mcnemar's Test P-Value : 7.34e-06
##
##             Sensitivity : 0.64286
##             Specificity : 0.62222
##          Pos Pred Value : 0.20930
##          Neg Pred Value : 0.91803
##              Prevalence : 0.13462
##          Detection Rate : 0.08654
##    Detection Prevalence : 0.41346
##       Balanced Accuracy : 0.63254
##
##        'Positive' Class : Down
##
```

Overall fraction of correct predictions for the held out data is accuracy is 0.625

e] Repeat (d) using linear discriminant analysis (LDA).

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

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction Down Up
##       Down    9 34
##       Up      5 56
##
##                Accuracy : 0.625
##                  95% CI : (0.5247, 0.718)
##     No Information Rate : 0.8654
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.1414
##  Mcnemar's Test P-Value : 7.34e-06
```

```
##
##             Sensitivity : 0.64286
##             Specificity : 0.62222
##          Pos Pred Value : 0.20930
##          Neg Pred Value : 0.91803
##              Prevalence : 0.13462
##          Detection Rate : 0.08654
##    Detection Prevalence : 0.41346
##       Balanced Accuracy : 0.63254
##
##        'Positive' Class : Down
##
```

Overall fraction of correct predictions for the held out data is accuracy is 0.625

f] Repeat (d) using quadratic discriminant analysis (QDA).

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

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

Overall fraction of correct predictions for the held out data is accuracy is 0.5865

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

```
##       knn.pred
##        Down Up
##   Down   21 22
##   Up     30 31
```

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

Overall fraction of correct predictions for the held out data is accuracy is 0.5865

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

The models from letter d and e, respectively Logistic Regression and LDA