Logistic regression and LDA

Vinoth Aryan Nagabosshanam

May 24, 2018

logistic regression model

building a model in a simple stack market data set

```
library(ISLR)
```

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

```
#View(Smarket)
colnames(Smarket)
```

```
## [1] "Year"      "Lag1"      "Lag2"      "Lag3"      "Lag4"      "Lag5"
## [7] "Volume"    "Today"     "Direction"
```

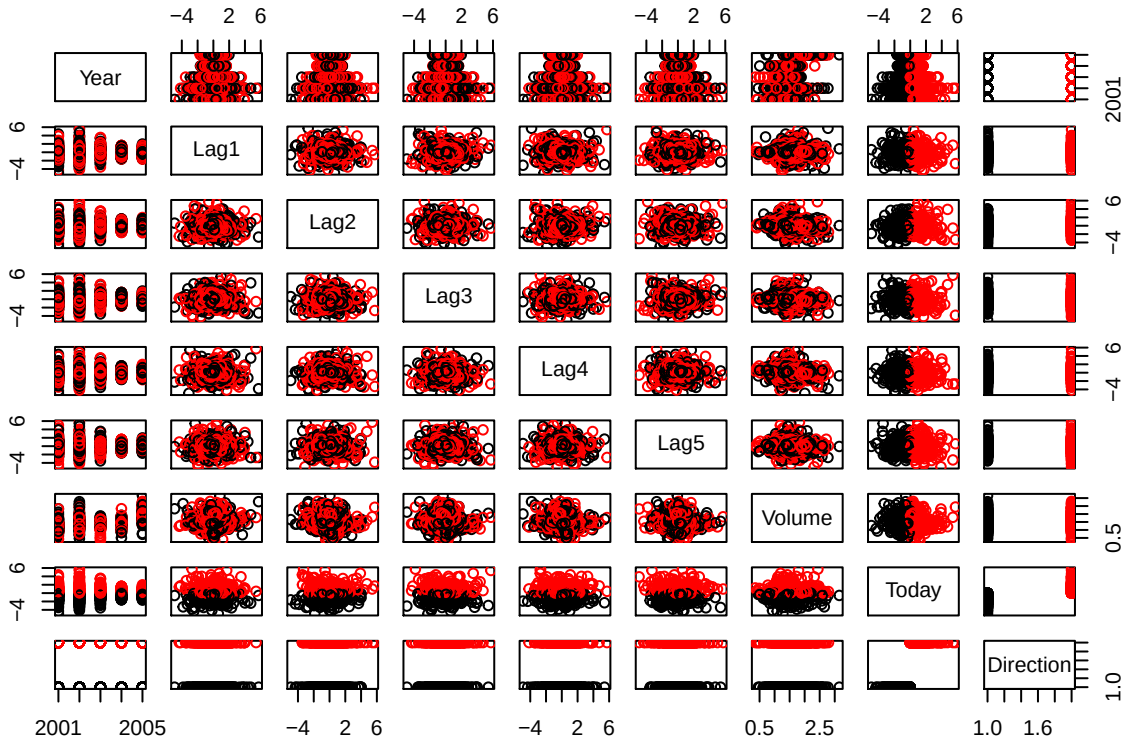
```
#str(Smarket)
summary(Smarket)
```

```
##       Year           Lag1                Lag2
##  Min.   :2001   Min.   :-4.922000   Min.   :-4.922000
##  1st Qu.:2002   1st Qu.:-0.639500   1st Qu.:-0.639500
##  Median :2003   Median : 0.039000   Median : 0.039000
##  Mean   :2003   Mean   : 0.003834   Mean   : 0.003919
##  3rd Qu.:2004   3rd Qu.: 0.596750   3rd Qu.: 0.596750
##  Max.   :2005   Max.   : 5.733000   Max.   : 5.733000
##       Lag3                Lag4                Lag5
##  Min.   :-4.922000   Min.   :-4.922000   Min.   :-4.92200
##  1st Qu.:-0.640000   1st Qu.:-0.640000   1st Qu.:-0.64000
##  Median : 0.038500   Median : 0.038500   Median : 0.03850
##  Mean   : 0.001716   Mean   : 0.001636   Mean   : 0.00561
##  3rd Qu.: 0.596750   3rd Qu.: 0.596750   3rd Qu.: 0.59700
##  Max.   : 5.733000   Max.   : 5.733000   Max.   : 5.73300
##      Volume           Today           Direction
##  Min.   :0.3561   Min.   :-4.922000   Down:602
##  1st Qu.:1.2574   1st Qu.:-0.639500   Up  :648
##  Median :1.4229   Median : 0.038500
##  Mean   :1.4783   Mean   : 0.003138
##  3rd Qu.:1.6417   3rd Qu.: 0.596750
##  Max.   :3.1525   Max.   : 5.733000
```

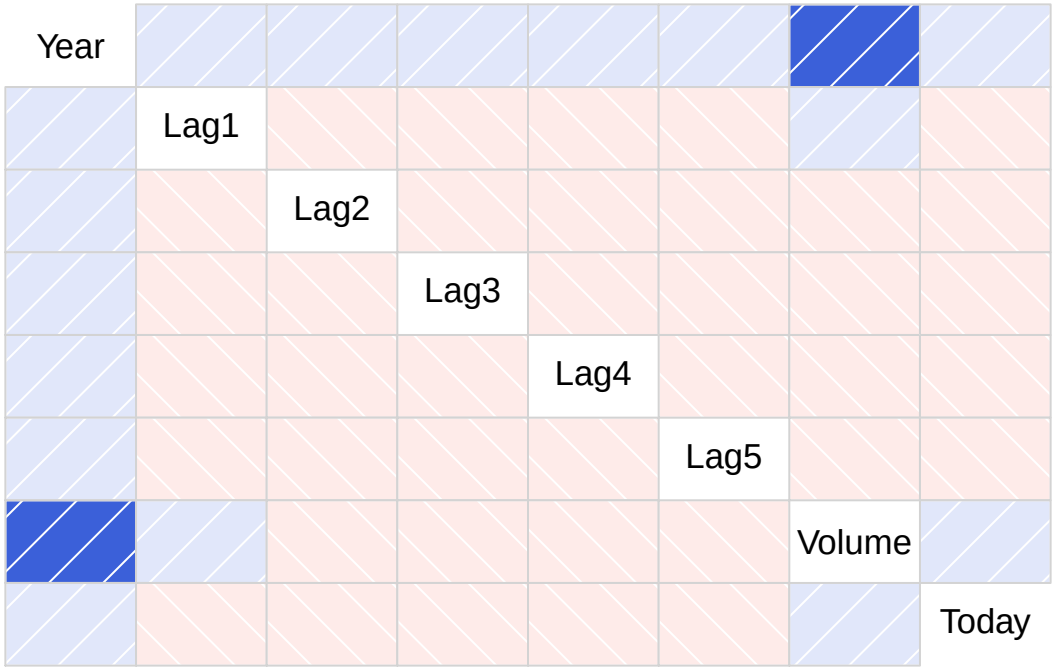
to check the corelation between the varience and we use the pair

```
pairs(Smarket,col=Smarket$Direction)
library(corrgram)
```

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




```
corrgram(Smarket)
```



we are bulid the logistic regression model where y is direction x is other variabale as input data

```
model_1<-glm(Direction~.-Year -Today,data=Smarket,family = binomial)
summary(model_1)
```

```
##
## Call:
## glm(formula = Direction ~ . - Year - Today, family = binomial,
##     data = Smarket)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.446  -1.203   1.065   1.145   1.326
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.126000   0.240736  -0.523    0.601
## Lag1        -0.073074   0.050167  -1.457    0.145
## Lag2        -0.042301   0.050086  -0.845    0.398
## Lag3         0.011085   0.049939   0.222    0.824
## Lag4         0.009359   0.049974   0.187    0.851
## Lag5         0.010313   0.049511   0.208    0.835
```

```
## Volume       0.135441   0.158360   0.855    0.392
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1731.2  on 1249  degrees of freedom
## Residual deviance: 1727.6  on 1243  degrees of freedom
## AIC: 1741.6
##
## Number of Fisher Scoring iterations: 3
```

```
exp(coef(model_1))
```

```
## (Intercept)        Lag1        Lag2        Lag3        Lag4        Lag5
##   0.8816146   0.9295323   0.9585809   1.0111468   1.0094029   1.0103664
##      Volume
##   1.1450412
```

```
prob <- predict(model_1,type=c("response"),Smarket)
#prob
confusion<-table(prob>0.5,Smarket$Direction)
confusion
```

```
##
##         Down  Up
##   FALSE  145 141
##   TRUE   457 507
```

```
# check the accuracy of the model
Accuracy<-sum(diag(confusion)/sum(confusion))
Accuracy
```

```
## [1] 0.5216
```

Make training and test set

```
train = Smarket$Year<2005

model_t<-glm(Direction~.-Year -Today,data=Smarket,family = binomial,subset=train)
summary(model_t)
```

```
##
## Call:
## glm(formula = Direction ~ . - Year - Today, family = binomial,
##     data = Smarket, subset = train)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.302  -1.190   1.079   1.160   1.350
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.191213   0.333690   0.573    0.567
## Lag1        -0.054178   0.051785  -1.046    0.295
## Lag2        -0.045805   0.051797  -0.884    0.377
## Lag3         0.007200   0.051644   0.139    0.889
```

```
## Lag4         0.006441   0.051706   0.125    0.901
## Lag5        -0.004223   0.051138  -0.083    0.934
## Volume      -0.116257   0.239618  -0.485    0.628
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1383.3  on 997  degrees of freedom
## Residual deviance: 1381.1  on 991  degrees of freedom
## AIC: 1395.1
##
## Number of Fisher Scoring iterations: 3
```

```
exp(coef(model_t))
```

```
## (Intercept)        Lag1        Lag2        Lag3        Lag4        Lag5
##   1.2107168   0.9472632   0.9552279   1.0072261   1.0064617   0.9957862
##      Volume
##   0.8902464
```

```
probs <- predict(model_t,type=c("response"),newdata=Smarket[!train,])
#prob
year_2005=Smarket$Direction[!train]
confusi<-table(probs>0.5,year_2005)
confusi
```

```
##        year_2005
##         Down Up
##   FALSE   77 97
##   TRUE    34 44
```

```
#find the accuarcy
accura<-sum(diag(confusi)/sum(confusi))
accura
```

```
## [1] 0.4801587
```

Fit smaller model

```
glm_fit=glm(Direction~Lag1+Lag2,
            data=Smarket,family=binomial, subset=train)
mprobs=predict(glm_fit,newdata=Smarket[!train,],type="response")

confus<-table(mprobs>0.5,year_2005)
confus
```

```
##        year_2005
##         Down  Up
##   FALSE   35  35
##   TRUE    76 106
```

```
#find the accuarcy
accur<-sum(diag(confus)/sum(confus))
accur
```

```
## [1] 0.5595238
```

same data set apply on Linear Discriminant Analysis

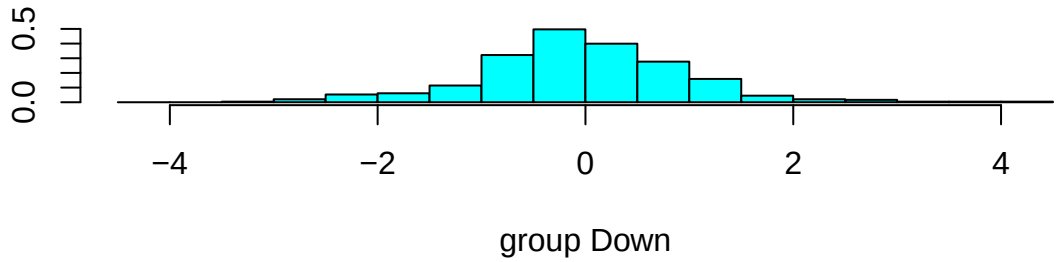
```
library(MASS)
```

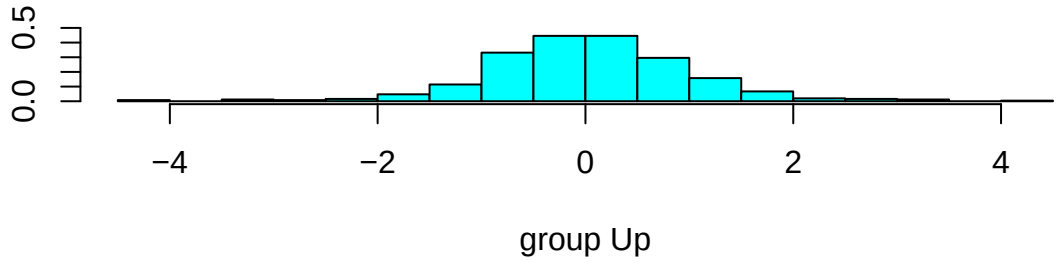
```
## Warning: package 'MASS' was built under R version 3.3.3
```

```
lda_model=lda(Direction~Lag1+Lag2,data=Smarket, subset=Year<2005)
lda_model
```

```
## Call:
## lda(Direction ~ Lag1 + Lag2, data = Smarket, subset = Year <
##     2005)
##
## Prior probabilities of groups:
##     Down       Up
## 0.491984 0.508016
##
## Group means:
##             Lag1        Lag2
## Down  0.04279022  0.03389409
## Up   -0.03954635 -0.03132544
##
## Coefficients of linear discriminants:
##             LD1
## Lag1 -0.6420190
## Lag2 -0.5135293
```

```
plot(lda_model)
```






```
Smarket.2005=subset(Smarket,Year==2005)
lda_pred=predict(lda_model,Smarket.2005)
class(lda_pred)
```

```
## [1] "list"
```

```
data.frame(lda_pred)[1:5,]
```

```
##      class posterior.Down posterior.Up         LD1
## 999     Up      0.4901792    0.5098208  0.08293096
## 1000    Up      0.4792185    0.5207815  0.59114102
## 1001    Up      0.4668185    0.5331815  1.16723063
## 1002    Up      0.4740011    0.5259989  0.83335022
## 1003    Up      0.4927877    0.5072123 -0.03792892
```

```
table(lda_pred$class,Smarket.2005$Direction)
```

```
##
##        Down  Up
##   Down   35  35
##   Up     76 106
```

```
mean(lda_pred$class==Smarket.2005$Direction)
```

```
## [1] 0.5595238
```