Assignment 7

Sonali Singh

# Lab-Logistic Regression

# Knowledge Mining: Logistic Regression  
# File: Lab\_logisticregression01.R  
# Theme: Logistic regression  
# Adapted from ISLR Chapter 4 Lab  
  
# Load ISLR library  
  
require(ISLR)

## Loading required package: ISLR

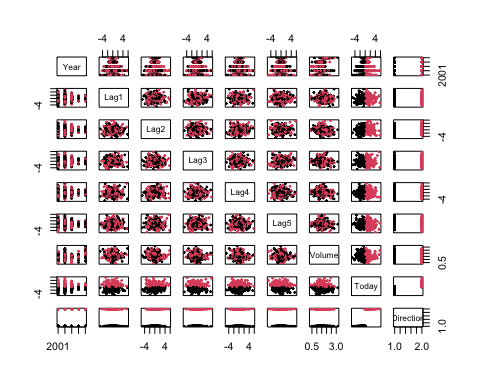
# Check dataset Smarket  
?Smarket  
names(Smarket)

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

summary(Smarket)

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

# Create a dataframe for data browsing  
sm=Smarket  
  
# Bivariate Plot of inter-lag correlations  
pairs(Smarket,col=Smarket$Direction,cex=.5, pch=20)



# Logistic regression  
glm.fit=glm(Direction~Lag1+Lag2+Lag3+Lag4+Lag5+Volume,  
 data=Smarket,family=binomial)  
summary(glm.fit)

##   
## Call:  
## glm(formula = Direction ~ Lag1 + Lag2 + Lag3 + Lag4 + Lag5 +   
## Volume, family = binomial, data = Smarket)  
##   
## 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

glm.probs=predict(glm.fit,type="response")   
glm.probs[1:5]

## 1 2 3 4 5   
## 0.5070841 0.4814679 0.4811388 0.5152224 0.5107812

glm.pred=ifelse(glm.probs>0.5,"Up","Down")  
attach(Smarket)  
table(glm.pred,Direction)

## Direction  
## glm.pred Down Up  
## Down 145 141  
## Up 457 507

mean(glm.pred==Direction)

## [1] 0.5216

# Make training and test set for prediction  
train = Year<2005  
glm.fit=glm(Direction~Lag1+Lag2+Lag3+Lag4+Lag5+Volume,  
 data=Smarket,family=binomial, subset=train)  
glm.probs=predict(glm.fit,newdata=Smarket[!train,],type="response")   
glm.pred=ifelse(glm.probs >0.5,"Up","Down")  
Direction.2005=Smarket$Direction[!train]  
table(glm.pred,Direction.2005)

## Direction.2005  
## glm.pred Down Up  
## Down 77 97  
## Up 34 44

mean(glm.pred==Direction.2005)

## [1] 0.4801587

#Fit smaller model  
glm.fit=glm(Direction~Lag1+Lag2,  
 data=Smarket,family=binomial, subset=train)  
glm.probs=predict(glm.fit,newdata=Smarket[!train,],type="response")   
glm.pred=ifelse(glm.probs >0.5,"Up","Down")  
table(glm.pred,Direction.2005)

## Direction.2005  
## glm.pred Down Up  
## Down 35 35  
## Up 76 106

mean(glm.pred==Direction.2005)

## [1] 0.5595238

# Check accuracy rate  
106/(76+106)

## [1] 0.5824176

# Can you interpret the results?

### Interpretation

1. **Data Summary and Structure**: The **Smarket** dataset contains daily percentage returns for the S&P 500 stock index from 2001 to 2005, characterized by features like **Lag1** to **Lag5**, **Volume**, and the response variable **Direction** (indicating whether the market went up or down).
2. **Pairwise Correlations**: The **pairs** plot visually assesses correlations and distributions of the variables against each other. Colored by the direction of market movement, it potentially identifies patterns or relationships critical for predictive modeling.
3. **Logistic Regression Model Development**:
   * **Initial Model**: A logistic regression model is fit with directional movements as a function of lags and volume. The coefficients are estimated, but none of the predictors appear statistically significant (p-values > 0.05), suggesting they may not reliably predict market direction.
   * **Performance**: The classification accuracy of the initial model on the dataset used for fitting is approximately 52.16%, which barely exceeds a random guess in a balanced binary classification.
4. **Model Validation**:
   * **Setup**: The data is split into a training set (before 2005) and a test set (2005 data).
   * **Refitted Model Performance on 2005 Data**: When the full model is applied to 2005 data, the accuracy is about 48.02%, indicating poor generalization, possibly due to overfitting or non-predictive lag variables.
   * **Simplified Model**: A simplified model using only **Lag1** and **Lag2** improves accuracy on the 2005 data to 55.95%. This suggests reducing complexity may help in avoiding overfitting and enhancing the model’s predictive accuracy.
5. **Accuracy Rate Calculation**: The final calculation, **106/(76+106) = 58.24%**, suggests a slightly better performance of the simplified model, specifically in the true positive predictions within the 2005 subset.

The overall effectiveness of lagged returns and volume in predicting market direction appears limited. The fluctuating accuracy rates highlight the challenges of financial time series forecasting, where past performance may not always predict future movements effectively.

## **Review ISLR Chapter 4 and Answers**

**2a. What is/are the requirement(s) of LDA?** Linear Discriminant Analysis (LDA) requires that the predictors are normally distributed within each class, and that the classes have similar covariance matrices.

**2b. How LDA is different from Logistic Regression?** LDA assumes that the predictors are normally distributed and the classes have the same covariance structure. Logistic regression does not make these assumptions and instead uses a logistic function to model the probability that an observation belongs to a particular class.

**2c. What is ROC?** ROC (Receiver Operating Characteristic) is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied. It is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings.

**2d. What is sensitivity and specificity? Which is more important in your opinion?** Sensitivity (or true positive rate) is the measure of the proportion of actual positives correctly identified. Specificity (or true negative rate) is the measure of the proportion of actual negatives correctly identified. The importance of each depends on the context; sensitivity might be prioritized in medical testing to avoid missing a diagnosis, while specificity might be emphasized in spam detection to avoid marking legitimate emails as spam.

**2e. From the following chart, for the purpose of prediction, which is more critical?** It depends on the consequences of false positives versus false negatives. For instance, in cancer screening, a false negative (a missed diagnosis) could be more critical than a false positive (an unnecessary scare).

### **Question 3-Calculate the prediction error from the following:**

To calculate the prediction error from the provided table:

* **True Negatives (TN)**: 9,644
* **True Positives (TP)**: 81
* **False Negatives (FN)**: 252
* **False Positives (FP)**: 23

The prediction error can be calculated as the sum of the instances where the predictions were incorrect over the total number of instances.

to compute this value:

So, the prediction error is 0.0275, or 2.75%.