ex5

Christopher Huong

2023-07-30

```
library(ISLR)
library(caret)
library(tidyverse)
library(psych)
library(pls)
```

Predicting "Direction" with logistic regression

```
data(Weekly)

str(Weekly)
```

```
## 'data.frame':    1089 obs. of  9 variables:
##  $ Year     : num  1990 1990 1990 1990 1990 1990 1990 1990 1990 1990 ...
##  $ Lag1     : num  0.816 -0.27 -2.576 3.514 0.712 ...
##  $ Lag2     : num  1.572 0.816 -0.27 -2.576 3.514 ...
##  $ Lag3     : num  -3.936 1.572 0.816 -0.27 -2.576 ...
##  $ Lag4     : num  -0.229 -3.936 1.572 0.816 -0.27 ...
##  $ Lag5     : num  -3.484 -0.229 -3.936 1.572 0.816 ...
##  $ Volume   : num  0.155 0.149 0.16 0.162 0.154 ...
##  $ Today    : num  -0.27 -2.576 3.514 0.712 1.178 ...
##  $ Direction: Factor w/ 2 levels "Down","Up": 1 1 2 2 2 1 2 2 2 1 ...
```

```
glimpse(Weekly)
```

```
## Rows: 1,089
## Columns: 9
## $ Year      <dbl> 1990, 1990, 1990, 1990, 1990, 1990, 1990, 1990, 1990, 1990, ~
## $ Lag1      <dbl> 0.816, -0.270, -2.576, 3.514, 0.712, 1.178, -1.372, 0.807, 0~
## $ Lag2      <dbl> 1.572, 0.816, -0.270, -2.576, 3.514, 0.712, 1.178, -1.372, 0~
## $ Lag3      <dbl> -3.936, 1.572, 0.816, -0.270, -2.576, 3.514, 0.712, 1.178, -~
## $ Lag4      <dbl> -0.229, -3.936, 1.572, 0.816, -0.270, -2.576, 3.514, 0.712, ~
## $ Lag5      <dbl> -3.484, -0.229, -3.936, 1.572, 0.816, -0.270, -2.576, 3.514,~
## $ Volume    <dbl> 0.1549760, 0.1485740, 0.1598375, 0.1616300, 0.1537280, 0.154~
## $ Today     <dbl> -0.270, -2.576, 3.514, 0.712, 1.178, -1.372, 0.807, 0.041, 1~
## $ Direction <fct> Down, Down, Up, Up, Up, Down, Up, Up, Up, Down, Down, Up, Up~
```

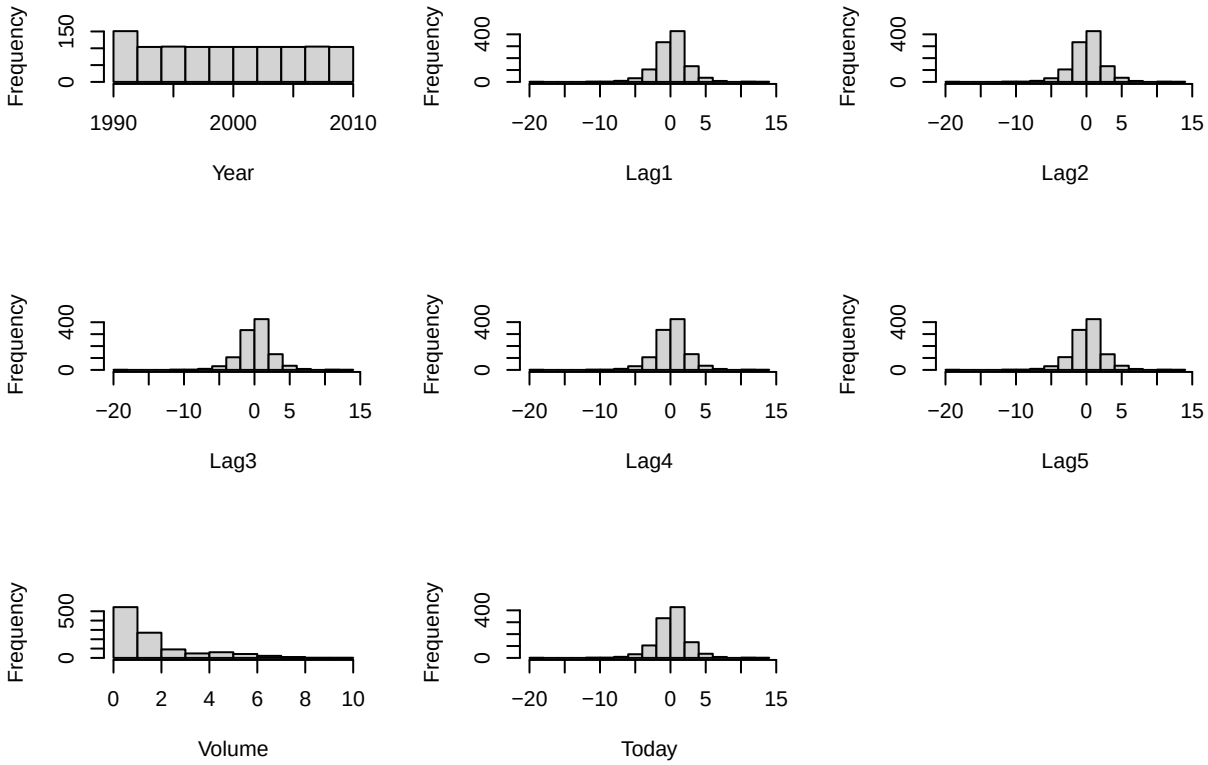
```
describe(Weekly)[,c(3,4,5,8,9,11,12)]
```

```
##                mean   sd  median     min     max  skew kurtosis
## Year       2000.05 6.03 2000.00 1990.00 2010.00  0.00    -1.21
## Lag1          0.15 2.36    0.24  -18.20   12.03 -0.48     5.67
## Lag2          0.15 2.36    0.24  -18.20   12.03 -0.48     5.67
## Lag3          0.15 2.36    0.24  -18.20   12.03 -0.48     5.62
## Lag4          0.15 2.36    0.24  -18.20   12.03 -0.48     5.63
## Lag5          0.14 2.36    0.23  -18.20   12.03 -0.47     5.61
## Volume        1.57 1.69    1.00    0.09    9.33  1.62     2.06
## Today         0.15 2.36    0.24  -18.20   12.03 -0.48     5.67
## Direction*    1.56 0.50    2.00    1.00    2.00 -0.22    -1.95
```

```r
table(Weekly$Year)
```

```
##
## 1990 1991 1992 1993 1994 1995 1996 1997 1998 1999 2000 2001 2002 2003 2004 2005
##   47   52   52   52   52   52   53   52   52   52   52   52   52   52   52   52
## 2006 2007 2008 2009 2010
##   52   53   52   52   52
```

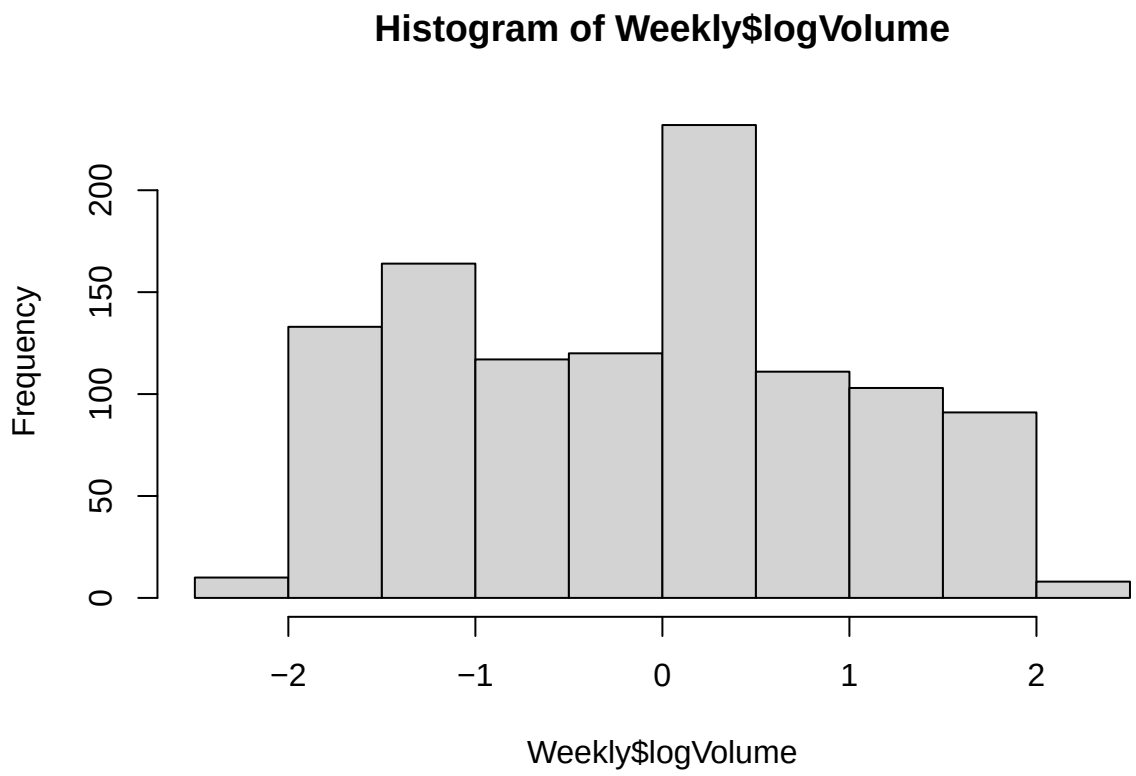
```r
vars_list <- as.list(colnames(select(Weekly,-Direction)))

par(mfrow=c(3,3))
for(i in vars_list){hist(select(Weekly,-Direction)[,i],xlab=i,main="")}
```



The year ranges from 1990 to 2010 with a roughly uniform distribution. All the Lag variables have roughly

equivalent distributions (mean, median, range, sd) Only the volume variable has significant skewness. A log transformation may be indicated. Most variables are relatively kurtotic (wider distribution).

```
Weekly$logVolume <- log(Weekly$Volume)
hist(Weekly$logVolume)
```

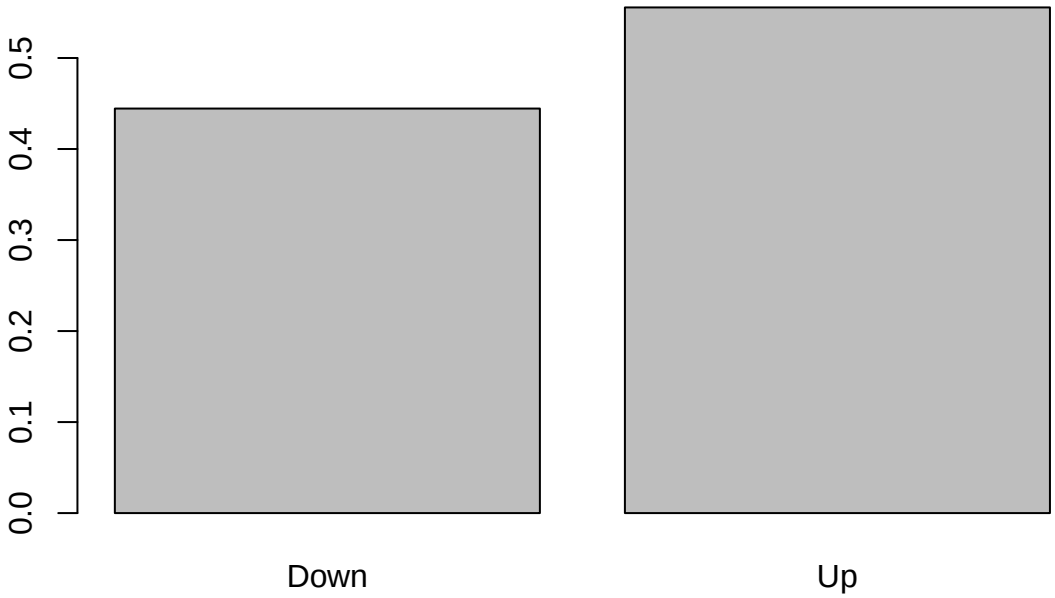



```
skew(Weekly$logVolume)
```

```
## [1] 0.05196872
```

Better

```
barplot(prop.table(table(Weekly$Direction)))
```



The response variable is binary and roughly equally distributed.

```
mod1 <- glm(Direction ~ Lag1+Lag2+Lag3+Lag4+Lag5+logVolume,
            data = Weekly,
            family = "binomial")

summary(mod1)
```

```
##
## Call:
## glm(formula = Direction ~ Lag1 + Lag2 + Lag3 + Lag4 + Lag5 +
##     logVolume, family = "binomial", data = Weekly)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6922  -1.2600   0.9928   1.0847   1.4665
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.22562    0.06224   3.625 0.000289 ***
## Lag1        -0.04127    0.02637  -1.565 0.117578
## Lag2         0.05834    0.02679   2.178 0.029433 *
## Lag3        -0.01607    0.02663  -0.603 0.546213
## Lag4        -0.02790    0.02643  -1.055 0.291218
## Lag5        -0.01457    0.02636  -0.553 0.580433
## logVolume   -0.05133    0.05607  -0.915 0.359988
```

```
## ---
## 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: 1485.9  on 1082  degrees of freedom
## AIC: 1499.9
##
## Number of Fisher Scoring iterations: 4
```

Only the Lag2 variable is a statistically significant predictor of Direction ($p<0.05$)

```
set.seed(123)
ctrl <- trainControl(method = "LGOCV",
                     summaryFunction = twoClassSummary,
                     classProbs = TRUE,
                     savePredictions = T)


predictors <- select(Weekly,c(Lag1,Lag2,Lag3,Lag4,Lag5,logVolume))

mod2 <- train(x=predictors, y=Weekly$Direction,
                     method = "glm",
                     metric = "ROC",
                     trControl = ctrl)

mod2
```

```
## Generalized Linear Model
##
## 1089 samples
##    6 predictor
##    2 classes: 'Down', 'Up'
##
## No pre-processing
## Resampling: Repeated Train/Test Splits Estimated (25 reps, 75%)
## Summary of sample sizes: 817, 817, 817, 817, 817, 817, ...
## Resampling results:
##
##   ROC        Sens       Spec
##   0.5296962  0.1252893  0.8908609
```

```
confusionMatrix(data = mod2$pred$pred,
 reference = mod2$pred$obs)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction Down   Up
##       Down  379  412
##       Up   2646 3363
##
```

```
##                Accuracy : 0.5503
##                  95% CI : (0.5384, 0.5622)
##     No Information Rate : 0.5551
##     P-Value [Acc > NIR] : 0.7932
##
##                   Kappa : 0.0174
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.12529
##             Specificity : 0.89086
##          Pos Pred Value : 0.47914
##          Neg Pred Value : 0.55966
##              Prevalence : 0.44485
##          Detection Rate : 0.05574
##    Detection Prevalence : 0.11632
##       Balanced Accuracy : 0.50808
##
##        'Positive' Class : Down
##
```

The logistic regression shows low sensitivity (lots of false negatives / low true positives) and high specificity (high true negatives / low false positives). The accuracy is 0.5503

```
set.seed(123)
training <- Weekly %>%
  filter(Year %in% 1990:2008) %>%
  select(c(Lag1,Lag2,Lag3,Lag4,Lag5,logVolume, Direction))

testing <- Weekly %>%
  filter(Year %in% 2009:2010) %>%
  select(c(Lag1,Lag2,Lag3,Lag4,Lag5,logVolume, Direction))


mod3 <- train(Direction ~ Lag1 + Lag2 + Lag3 + Lag4 + Lag5 + logVolume,
              data = training,
              method = "glm",
              trControl = ctrl,
              family = binomial)
```

```
## Warning in train.default(x, y, weights = w, ...): The metric "Accuracy" was not
## in the result set. ROC will be used instead.
```

```
mod3
```

```
## Generalized Linear Model
##
## 985 samples
##   6 predictor
##   2 classes: 'Down', 'Up'
##
## No pre-processing
## Resampling: Repeated Train/Test Splits Estimated (25 reps, 75%)
```

```
## Summary of sample sizes: 739, 739, 739, 739, 739, 739, ...
## Resampling results:
##
##   ROC        Sens       Spec
##   0.5368422  0.1898182  0.8444118
```

```
predictions <- predict(mod3, newdata = testing)

confusionMatrix(predictions, testing$Direction)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction Down Up
##       Down   23 31
##       Up     20 30
##
##                Accuracy : 0.5096
##                  95% CI : (0.4097, 0.609)
##     No Information Rate : 0.5865
##     P-Value [Acc > NIR] : 0.9540
##
##                   Kappa : 0.0257
##
##  Mcnemar's Test P-Value : 0.1614
##
##             Sensitivity : 0.5349
##             Specificity : 0.4918
##          Pos Pred Value : 0.4259
##          Neg Pred Value : 0.6000
##              Prevalence : 0.4135
##          Detection Rate : 0.2212
##    Detection Prevalence : 0.5192
##       Balanced Accuracy : 0.5133
##
##        'Positive' Class : Down
##
```

The prediction model trained on years 1990-2008 has mediocre sensitivity and specificity on the testing data of 2009-2010. The accuracy is 0.5096

Now with LDA

```
set.seed(123)
mod4 <- train(Direction ~ Lag1 + Lag2 + Lag3 + Lag4 + Lag5 + logVolume,
              data = training,
              method = "lda",
              trControl = ctrl,
              family = binomial)
```

```
## Warning in train.default(x, y, weights = w, ...): The metric "Accuracy" was not
## in the result set. ROC will be used instead.
```

```
mod4
```

```
## Linear Discriminant Analysis
##
## 985 samples
##   6 predictor
##   2 classes: 'Down', 'Up'
##
## No pre-processing
## Resampling: Repeated Train/Test Splits Estimated (25 reps, 75%)
## Summary of sample sizes: 739, 739, 739, 739, 739, 739, ...
## Resampling results:
##
##   ROC       Sens       Spec
##   0.537016  0.1818182  0.8482353
```

```
predictions2 <- predict(mod4, newdata = testing)

confusionMatrix(predictions2, testing$Direction)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction Down Up
##       Down   23 30
##       Up     20 31
##
##                Accuracy : 0.5192
##                  95% CI : (0.4191, 0.6183)
##     No Information Rate : 0.5865
##     P-Value [Acc > NIR] : 0.9316
##
##                   Kappa : 0.0417
##
##  Mcnemar's Test P-Value : 0.2031
##
##             Sensitivity : 0.5349
##             Specificity : 0.5082
##          Pos Pred Value : 0.4340
##          Neg Pred Value : 0.6078
##              Prevalence : 0.4135
##          Detection Rate : 0.2212
##    Detection Prevalence : 0.5096
##       Balanced Accuracy : 0.5215
##
##        'Positive' Class : Down
##
```

The LDA prediction model has mediocre sensitivity and specificity, and an accuracy of 0.5192, which is superior to the logistic regression model.

Now with Partial least squares discriminant analysis

```r
set.seed(123)

mod5 <- plsda(select(training, -Direction), y=training$Direction, ncomp = 3)


pred3 <- predict(mod5, newdata = select(testing,-Direction))


confusionMatrix(pred3, testing$Direction)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction Down Up
##       Down   22 31
##       Up     21 30
##
##                Accuracy : 0.5
##                  95% CI : (0.4003, 0.5997)
##     No Information Rate : 0.5865
##     P-Value [Acc > NIR] : 0.970
##
##                   Kappa : 0.0033
##
##  Mcnemar's Test P-Value : 0.212
##
##             Sensitivity : 0.5116
##             Specificity : 0.4918
##          Pos Pred Value : 0.4151
##          Neg Pred Value : 0.5882
##              Prevalence : 0.4135
##          Detection Rate : 0.2115
##    Detection Prevalence : 0.5096
##       Balanced Accuracy : 0.5017
##
##        'Positive' Class : Down
##
```

The PLSDA has mediocre sensitivity and specificity, and correctly classified 0.5 of cases

Now with nearest shrunken centroids

```r
set.seed(123)
mod6 <- train(Direction ~ Lag1 + Lag2 + Lag3 + Lag4 + Lag5 + logVolume,
              data = training,
              method = "pam",
              trControl = ctrl,
              tuneGrid = data.frame(.threshold = 0:25))
```

```
## Warning in train.default(x, y, weights = w, ...): The metric "Accuracy" was not
## in the result set. ROC will be used instead.
```

```
## 1111111111111111111111111
```

```
mod6
```

```
## Nearest Shrunken Centroids
##
## 985 samples
##   6 predictor
##   2 classes: 'Down', 'Up'
##
## No pre-processing
## Resampling: Repeated Train/Test Splits Estimated (25 reps, 75%)
## Summary of sample sizes: 739, 739, 739, 739, 739, 739, ...
## Resampling results across tuning parameters:
##
##   threshold  ROC        Sens         Spec
##    0         0.5354733  0.002909091  0.9955882
##    1         0.4997540  0.000000000  1.0000000
##    2         0.5000000  0.000000000  1.0000000
##    3         0.5000000  0.000000000  1.0000000
##    4         0.5000000  0.000000000  1.0000000
##    5         0.5000000  0.000000000  1.0000000
##    6         0.5000000  0.000000000  1.0000000
##    7         0.5000000  0.000000000  1.0000000
##    8         0.5000000  0.000000000  1.0000000
##    9         0.5000000  0.000000000  1.0000000
##   10         0.5000000  0.000000000  1.0000000
##   11         0.5000000  0.000000000  1.0000000
##   12         0.5000000  0.000000000  1.0000000
##   13         0.5000000  0.000000000  1.0000000
##   14         0.5000000  0.000000000  1.0000000
##   15         0.5000000  0.000000000  1.0000000
##   16         0.5000000  0.000000000  1.0000000
##   17         0.5000000  0.000000000  1.0000000
##   18         0.5000000  0.000000000  1.0000000
##   19         0.5000000  0.000000000  1.0000000
##   20         0.5000000  0.000000000  1.0000000
##   21         0.5000000  0.000000000  1.0000000
##   22         0.5000000  0.000000000  1.0000000
##   23         0.5000000  0.000000000  1.0000000
##   24         0.5000000  0.000000000  1.0000000
##   25         0.5000000  0.000000000  1.0000000
##
## ROC was used to select the optimal model using the largest value.
## The final value used for the model was threshold = 0.
```

```
predictions4 <- predict(mod6, newdata = testing)

confusionMatrix(predictions4, testing$Direction)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction Down Up
##       Down    0  0
```

```
##       Up     43 61
##
##                Accuracy : 0.5865
##                  95% CI : (0.4858, 0.6823)
##     No Information Rate : 0.5865
##     P-Value [Acc > NIR] : 0.5419
##
##                   Kappa : 0
##
##  Mcnemar's Test P-Value : 1.504e-10
##
##             Sensitivity : 0.0000
##             Specificity : 1.0000
##          Pos Pred Value :    NaN
##          Neg Pred Value : 0.5865
##              Prevalence : 0.4135
##          Detection Rate : 0.0000
##    Detection Prevalence : 0.0000
##       Balanced Accuracy : 0.5000
##
##        'Positive' Class : Down
##
```

Something may be wrong, as the model is classifying all predictions as Up, which is mostly accurate (0.5865) and thus performs better than the other models.