1. Produce some numerical and graphical summaries of the “Weekly” data. Do there appear to be any patterns ?

Ans - Year Lag1

Min. :1990 Min. :-18.1950

1st Qu.:1995 1st Qu.: -1.1540

Median :2000 Median : 0.2410

Mean :2000 Mean : 0.1506

3rd Qu.:2005 3rd Qu.: 1.4050

Max. :2010 Max. : 12.0260

Lag2 Lag3

Min. :-18.1950 Min. :-18.1950

1st Qu.: -1.1540 1st Qu.: -1.1580

Median : 0.2410 Median : 0.2410

Mean : 0.1511 Mean : 0.1472

3rd Qu.: 1.4090 3rd Qu.: 1.4090

Max. : 12.0260 Max. : 12.0260

Lag4 Lag5

Min. :-18.1950 Min. :-18.1950

1st Qu.: -1.1580 1st Qu.: -1.1660

Median : 0.2380 Median : 0.2340

Mean : 0.1458 Mean : 0.1399

3rd Qu.: 1.4090 3rd Qu.: 1.4050

Max. : 12.0260 Max. : 12.0260

Volume Today Direction

Min. :0.08747 Min. :-18.1950 Down:484

1st Qu.:0.33202 1st Qu.: -1.1540 Up :605

Median :1.00268 Median : 0.2410

Mean :1.57462 Mean : 0.1499

3rd Qu.:2.05373 3rd Qu.: 1.4050

Max. :9.32821 Max. : 12.0260

Year Lag1 Lag2

Year 1.00000000 -0.032289274 -0.03339001

Lag1 -0.03228927 1.000000000 -0.07485305

Lag2 -0.03339001 -0.074853051 1.00000000

Lag3 -0.03000649 0.058635682 -0.07572091

Lag4 -0.03112792 -0.071273876 0.05838153

Lag5 -0.03051910 -0.008183096 -0.07249948

Volume 0.84194162 -0.064951313 -0.08551314

Today -0.03245989 -0.075031842 0.05916672

Lag3 Lag4 Lag5

Year -0.03000649 -0.031127923 -0.030519101

Lag1 0.05863568 -0.071273876 -0.008183096

Lag2 -0.07572091 0.058381535 -0.072499482

Lag3 1.00000000 -0.075395865 0.060657175

Lag4 -0.07539587 1.000000000 -0.075675027

Lag5 0.06065717 -0.075675027 1.000000000

Volume -0.06928771 -0.061074617 -0.058517414

Today -0.07124364 -0.007825873 0.011012698

Volume Today

Year 0.84194162 -0.032459894

Lag1 -0.06495131 -0.075031842

Lag2 -0.08551314 0.059166717

Lag3 -0.06928771 -0.071243639

Lag4 -0.06107462 -0.007825873

Lag5 -0.05851741 0.011012698

Volume 1.00000000 -0.033077783

Today -0.03307778 1.000000000

Chart, histogram, scatter chart

Description automatically generated

As we can see above the correlations between the lag variables and the today’s returns are close to zero. But the only substantial correlation is between Year and Volume. And when we plot Volume, we see that it is increasing over time.

1. 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 appear to be statistically significant ? If so, which ones ?

Ans – As seen below It would seem that “Lag2” is the only predictor statistically significant as its p-value is less than 0.05.

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

1. 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.

Ans – From what we have below we may conclude that the percentage of correct predictions on the training data is (54+557)/1089 which will equal to 56.1065197%. We could also say that for weeks when the market goes up, the model is right 92.0661157% of the time (557/(48+557) .And for the weeks when the market goes down, the model is right only 11.1570248% of the time (54/(54+430)

Direction

pred.glm Down Up

Down 54 48

Up 430 557

1. 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 to 2010).

Ans - 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

> probs <- predict(fit.glm, type = "response")

> pred.glm <- rep("Down", length(probs))

> pred.glm[probs > 0.5] <- "Up"

> table(pred.glm, Direction)

Direction

pred.glm Down Up

Down 54 48

Up 430 557

> train <- (Year < 2009)

> Weekly.20092010 <- Weekly[!train, ]

> Direction.20092010 <- Direction[!train]

> fit.glm2 <- glm(Direction ~ Lag2, data = Weekly, family = binomial, subset = train)

> summary(fit.glm2)

Call:

glm(formula = Direction ~ Lag2, family = binomial, data = Weekly,

subset = train)

Deviance Residuals:

Min 1Q Median 3Q Max

-1.536 -1.264 1.021 1.091 1.368

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) 0.20326 0.06428 3.162 0.00157 \*\*

Lag2 0.05810 0.02870 2.024 0.04298 \*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 1354.7 on 984 degrees of freedom

Residual deviance: 1350.5 on 983 degrees of freedom

AIC: 1354.5

Number of Fisher Scoring iterations: 4

Direction.20092010

pred.glm2 Down Up

Down 9 5

Up 34 56

From what we can see up there we may conclude that the percentage of correct predictions on the test data is (9+56)/104 which will equal to 62.5%. But in other words 37.5% is the test error rate. We could also say that for weeks when the market goes up, the model is right 91.8032787% of the time (56/(56+5). For weeks when the market goes down, the model is right only 20.9302326% of the time (9/(9+34)

1. Repeat (d) using LDA.

Ans - Call:

lda(Direction ~ Lag2, data = Weekly, subset = train)

Prior probabilities of groups:

Down Up

0.4477157 0.5522843

Group means:

Lag2

Down -0.03568254

Up 0.26036581

Coefficients of linear discriminants:

LD1

Lag2 0.4414162

Direction.20092010

Down Up

Down 9 5

Up 34 56

From what we see up there we may conclude that the percentage of correct predictions on the test data is 62.5%. In other words 37.5% is the test error rate. We could also say that for weeks when the market goes up, the model is right 91.8032787% of the time. For weeks when the market goes down, the model is right only 20.9302326% of the time.

1. Repeat (d) using QDA.

Ans - Call:

qda(Direction ~ Lag2, data = Weekly, subset = train)

Prior probabilities of groups:

Down Up

0.4477157 0.5522843

Group means:

Lag2

Down -0.03568254

Up 0.26036581

Direction.20092010

Down Up

Down 0 0

Up 43 61

From what we got from using the QDA , we may conclude that the percentage of correct predictions on the test data is 58.6538462%. In other words 41.3461538% is the test error rate. We could also say that for weeks when the market goes up, the model is right 100% of the time. For weeks when the market goes down, the model is right only 0% of the time. We may note, that QDA achieves a correctness of 58.6538462% even though the model chooses “Up” the whole time !

1. Repeat (d) using KNN with K=1.

Ans - Direction.20092010

pred.knn Down Up

Down 21 30

Up 22 31

From this case, we may conclude that the percentage of correct predictions on the test data is 50%. In other words 50% is the test error rate. We could also say that for weeks when the market goes up, the model is right 50.8196721% of the time. For weeks when the market goes down, the model is right only 48.8372093% of the time.

1. Which of these methods appears to provide the best results on this data ?

Ans – I would say if we were tocompare the test error rates, we will see that logistic regression and LDA have the minimum error rates, followed by QDA and KNN.

1. Experiment with different combinations of predictors, including possible transformations and interactions, for each of the methods. Report the variables, method, and associated confusion matrix that appears to provide the best results on the held out data. Note that you should also experiment with values for K in the KNN classifier.

Ans - Direction.20092010

pred.glm3 Down Up

Down 1 1

Up 42 60

Here we have the Logistic regression with Lag2:Lag1 -> [1] 0.5865385

And here we have the LDA with Lag2 interaction with Lag1 -> [1] 0.5769231

Below we have the QDA with sqrt

Direction.20092010

Down Up

Down 12 13

Up 31 48

[1] 0.5769231

Here we have the KNN K =10

Direction.20092010

pred.knn2 Down Up

Down 17 18

Up 26 43

[1] 0.5769231

Here we have the KNN K =100

Direction.20092010

pred.knn3 Down Up

Down 9 12

Up 34 49

[1] 0.5576923

And in conclusion out of all these combinations we made so far , the original logistic regression and LDA have the best performance in terms of test error rates.

CHAPTER 5

2

1. Fit a logistic regression model that predicts “Direction” using “Lag1” and “Lag2”.

Ans - Call:

glm(formula = Direction ~ Lag1 + Lag2, family = "binomial", data = Weekly)

Deviance Residuals:

Min 1Q Median 3Q Max

-1.623 -1.261 1.001 1.083 1.506

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) 0.22122 0.06147 3.599 0.000319

Lag1 -0.03872 0.02622 -1.477 0.139672

Lag2 0.06025 0.02655 2.270 0.023232

(Intercept) \*\*\*

Lag1

Lag2 \*

---

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: 1488.2 on 1086 degrees of freedom

AIC: 1494.2

Number of Fisher Scoring iterations: 4

1. Fit a logistic regression model that predicts “Direction” using “Lag1” and “Lag2” using all but the first observation.

Ans - Call:

glm(formula = Direction ~ Lag1 + Lag2, family = "binomial", data = Weekly[-1,

])

Deviance Residuals:

Min 1Q Median 3Q Max

-1.6258 -1.2617 0.9999 1.0819 1.5071

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) 0.22324 0.06150 3.630 0.000283

Lag1 -0.03843 0.02622 -1.466 0.142683

Lag2 0.06085 0.02656 2.291 0.021971

(Intercept) \*\*\*

Lag1

Lag2 \*

---

Signif. codes:

0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 1494.6 on 1087 degrees of freedom

Residual deviance: 1486.5 on 1085 degrees of freedom

AIC: 1492.5

Number of Fisher Scoring iterations: 4

1. Use the model from (b) to predict the direction of the first observation. You can do this by predicting that the first observation will go up if P(direction = Up|Lag1,Lag2)>0.5Was this observation correctly classified ?

Ans - 1

TRUE

From this we may conclude that the prediction for the first observation is “Up”. This observation was not correctly classified as the true direction is “Down”.

1. Write a loop from i=1 to i=n, where n is the number of observations in the data set, that performs each of the following steps :
2. Fit a logistic regression model using all but the ith observation to predict “Direction” using “Lag1” and “Lag2”.
3. Compute the posterior probability of the market moving up for the ith observation.
4. Use the posterior probability for the ith observation in order to predict whether or not the market moves up.
5. Determine whether or not an error was made in predicting the direction for the ith observation. If an error was made, then indicate this as a 1, and otherwise indicate it as a 0.

[1] 1 1 0 1 0 1 0 0 0 1 1 0 1 0 1 0 1 0 1 0 0 0 1 1 1 1 1 1 0 1 1 1 1 0

[35] 1 0 0 0 1 0 1 0 0 1 0 1 1 1 0 1 0 0 0 1 0 0 1 1 0 0 0 0 1 0 1 1 0 0

[69] 1 0 1 1 0 0 0 1 0 1 1 0 0 1 1 0 1 1 0 0 1 0 0 1 1 1 0 0 0 0 0 1 0 1

[103] 1 0 0 1 0 1 0 0 1 1 0 0 1 0 0 1 0 0 1 1 1 1 0 0 0 1 0 1 0 1 1 0 0 0

[137] 1 1 1 0 0 0 1 0 0 0 0 0 0 1 1 1 0 1 0 0 1 1 0 1 0 0 1 1 0 0 1 0 0 1

[171] 0 0 1 1 1 0 1 0 1 0 0 0 0 0 0 0 0 1 1 0 1 0 1 0 1 0 1 0 0 1 0 0 1 0

[205] 0 1 0 1 0 1 1 1 0 0 1 1 0 1 0 0 1 1 0 0 0 1 1 1 0 1 0 1 0 1 0 0 0 1

[239] 1 0 1 0 1 0 1 0 1 0 1 1 0 1 0 0 1 0 0 1 0 0 0 0 0 1 0 0 0 1 0 0 1 0

[273] 0 0 1 0 0 1 0 0 1 0 0 1 0 1 1 0 0 0 0 0 1 0 1 0 0 1 0 0 0 1 0 0 1 1

[307] 0 0 1 0 0 0 0 1 0 1 1 0 0 1 0 1 0 1 1 0 0 0 1 0 1 0 0 1 1 1 1 0 1 0

[341] 0 1 0 0 0 1 0 1 0 1 0 0 0 0 0 1 1 0 0 1 0 0 1 0 0 0 1 1 0 1 1 1 1 1

[375] 0 0 0 1 0 0 0 0 0 0 1 0 1 1 0 0 1 1 0 0 0 0 0 1 0 0 1 1 1 0 1 0 1 0

[409] 1 1 1 0 1 0 0 0 0 0 0 0 0 0 0 1 0 1 0 1 0 1 0 0 1 0 1 0 0 0 0 0 1 1

[443] 1 1 0 1 1 0 1 0 1 1 0 1 0 0 1 0 0 1 1 0 0 0 0 1 1 0 0 1 0 1 0 0 0 1

[477] 0 0 1 0 0 0 1 1 1 0 1 0 0 0 1 0 1 1 1 0 0 0 0 1 1 1 0 1 1 0 1 0 0 0

[511] 1 0 1 0 0 0 1 0 1 1 0 0 1 1 0 0 0 1 1 0 1 0 1 1 1 1 1 0 0 0 1 0 0 0

[545] 1 1 0 1 0 0 0 1 1 1 1 1 1 0 1 0 1 0 0 1 0 0 1 1 1 0 0 0 1 1 1 1 1 1

[579] 1 1 1 1 0 1 0 0 1 0 0 1 0 1 0 0 1 0 0 1 0 1 1 0 1 1 1 0 1 0 1 0 1 0

[613] 0 0 1 0 1 0 1 0 1 1 0 1 1 0 1 0 1 0 1 1 1 1 0 1 1 0 0 0 1 1 1 1 0 1

[647] 1 1 0 1 0 0 0 1 1 1 1 1 1 0 1 0 0 1 0 0 0 1 1 0 1 0 1 1 1 1 0 0 0 1

[681] 1 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 1 0 0 1 0 1 1 0 0 0 0 1 0 1 0 1 0 1

[715] 0 0 1 1 0 0 0 0 0 0 0 0 0 1 0 0 1 0 0 1 1 1 0 1 1 0 1 1 1 1 0 0 0 1

[749] 1 1 1 1 1 0 1 0 0 0 0 0 0 1 0 1 1 1 0 0 0 1 0 0 1 0 0 0 1 1 1 0 0 0

[783] 1 0 0 1 1 1 0 0 1 0 1 0 1 0 0 1 0 0 1 0 0 0 0 0 1 0 1 1 0 0 1 1 0 1

[817] 1 1 0 0 0 0 0 1 1 0 0 1 0 0 1 0 1 0 0 0 1 1 0 1 1 0 1 0 1 0 1 1 0 0

[851] 1 1 1 0 1 1 0 0 0 1 0 1 0 1 0 1 0 0 0 0 0 1 0 0 1 1 0 0 1 0 1 0 1 1

[885] 0 1 0 1 1 0 0 0 1 0 0 0 0 0 0 0 1 0 1 0 1 0 0 0 1 1 1 1 1 0 1 1 0 0

[919] 0 0 0 1 0 0 1 0 0 0 0 1 0 1 1 1 0 0 1 1 0 1 1 1 1 0 1 0 1 0 1 0 1 0

[953] 1 0 0 1 1 1 1 1 0 1 0 0 0 1 1 1 0 1 1 1 1 0 0 0 0 1 1 0 0 0 0 1 0 0

[987] 1 1 1 0 0 1 1 1 0 1 0 0 0 0

[ reached getOption("max.print") -- omitted 89 entries ]

1. Take the average of the n numbers obtained in (d)iv in order to obtain the LOOCV estimate for the test error. Comment on the results.

Ans - [1] 0.4499541

We can see that for the LOOCV estimate for the test error rate is 44.9954086%.

CHAPTER 5 EXCERISE 9

1. Based on this data set, provide an estimate for the population mean of medv. Call this estimate μ.

Ans - μ= 22.5328063

#### Provide an estimate of the standard error of μ. Interpret this result.

Ans **-** The standard error is 0.4088611, found by the formula SEμ^=σ2n−−√) Now estimate the standard error of μusing the bootstrap. How does this compare to your answer from (b)?

Ans Our simulated standard error is SE = 0.4116822. It strongly agrees with the calculated standard error.

#### Based on your bootstrap estimate from (c), provide a 95 % confidence interval for the mean of medv. Compare it to the results obtained using t.test(Boston$medv)

Ans - lower

<dbl>

mean

<dbl>

upper

<dbl>

21.72591 22.53281 23.3397

One Sample t-test

data: Boston$medv

t = 55.111, df = 505, p-value < 2.2e-16

alternative hypothesis: true mean is not equal to 0

95 percent confidence interval:

21.72953 23.33608

sample estimates:

mean of x

22.53281

#### e) Based on this data set, provide an estimate, μ^med, for the median value of medv in the population.

Ans - The median is μ^med= 21.2

#### f) We now would like to estimate the standard error of μ^med. Unfortunately, there is no simple formula for computing the standard error of the median. Instead, estimate the standard error of the median using the bootstrap. Comment on your findings.

#### Ans - The simulated standard error of the median is SEmedian= 0.3793884

#### g) Based on this data set, provide an estimate for the tenth percentile of medv in Boston suburbs. Call this quantity μ^0.1.

#### Ans - The tenth percentile is 12.75.

#### h) Use the bootstrap to estimate the standard error of μ^0.1. Comment on your findings.

#### Ans - We find the se to be SEQ10 = 0.4905694