Homework 2

Xingyuan Chen

2/28/2019

7.

Suppose that we wish to predict whether a given stock will issue a dividend this year ("Yes" or "No") based on X, last year's percent profit. We examine a large number of companies and discover that the mean value of X for companies that issued a dividend was $\bar{X} = 10$, while the mean for those that didn't was $\bar{X} = 0$. In addition, the variance of X for these two sets of companies was $\hat{\sigma}^2 = 36$. Finally, 80% of companies issued dividends. Assuming that X follows a normal distribution, predict the probability that a company will issue a dividend this year given that its percentage profit was $X = 4$ last year.

Answer:
According to the question, we know that:

$$\mu(X|Yes) = 10, \mu(X|No) = 0, \sigma(X) = 6, P(Yes) = 0.8, P(No) = 0.2,$$

since:

$$P(X = x|Yes/No) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

so:

$$P(Yes|X = 4) = \frac{P(X = 4|Yes)P(Yes)}{P(X = 4|Yes)P(Yes) + P(X = 4|No)P(No)} = \frac{e^{-\frac{(4-10)^2}{2\times6^2}} \times 0.8}{e^{-\frac{(4-10)^2}{2\times6^2}} \times 0.8 + e^{-\frac{(4-0)^2}{2\times6^2}} \times 0.2} = 0.752$$

10.

This question should be answered using the `Weekly` data set, which is part of the `ISLR` package. This data is similar in nature to the `Smarket` data from this chapter's lab, except that it contains 1,089 weekly returns for 21 years, from the beginning of 1990 to the end of 2010.

(a) Produce some numerical and graphical summaries of the `Weekly` data. Do there appear to be any patterns?

```
summary(Weekly)
```

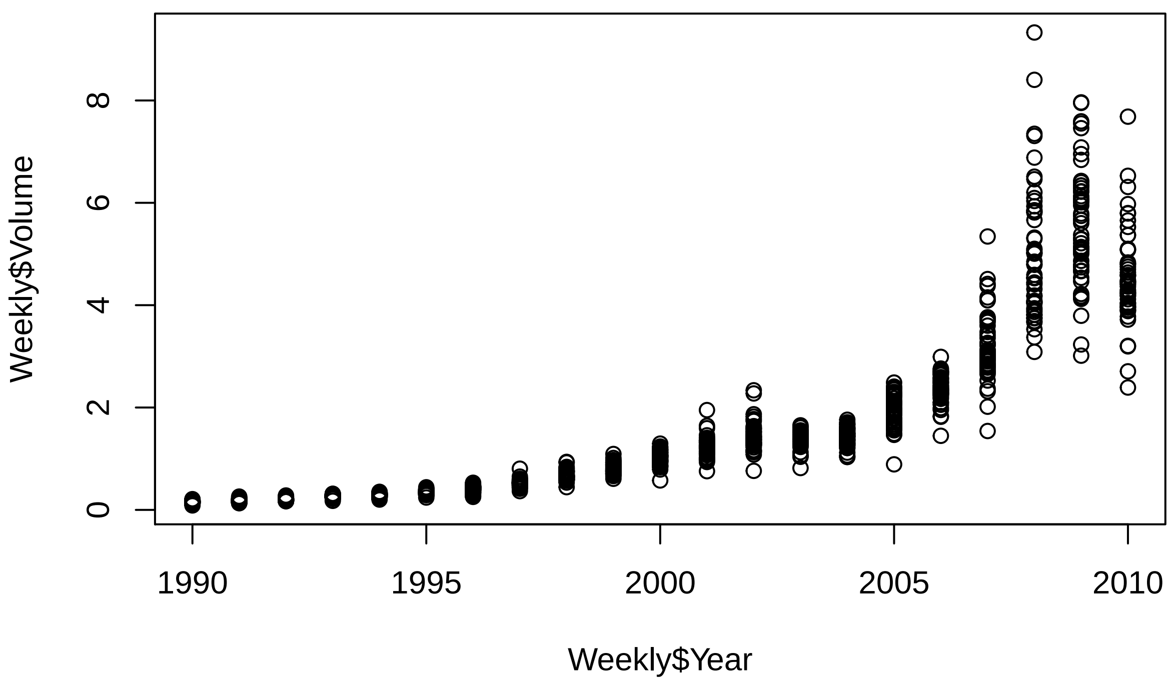
```
##       Year           Lag1               Lag2               Lag3
##  Min.   :1990   Min.   :-18.1950   Min.   :-18.1950   Min.   :-18.1950
##  1st Qu.:1995   1st Qu.: -1.1540   1st Qu.: -1.1540   1st Qu.: -1.1580
##  Median :2000   Median :  0.2410   Median :  0.2410   Median :  0.2410
##  Mean   :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
```

```r
cor(Weekly[,1:8])
```

```
##               Year         Lag1        Lag2        Lag3         Lag4
## Year    1.00000000 -0.032289274 -0.03339001 -0.03000649 -0.031127923
## Lag1   -0.03228927  1.000000000 -0.07485305  0.05863568 -0.071273876
## Lag2   -0.03339001 -0.074853051  1.00000000 -0.07572091  0.058381535
## Lag3   -0.03000649  0.058635682 -0.07572091  1.00000000 -0.075395865
## Lag4   -0.03112792 -0.071273876  0.05838153 -0.07539587  1.000000000
## Lag5   -0.03051910 -0.008183096 -0.07249948  0.06065717 -0.075675027
## Volume  0.84194162 -0.064951313 -0.08551314 -0.06928771 -0.061074617
## Today  -0.03245989 -0.075031842  0.05916672 -0.07124364 -0.007825873
##                Lag5      Volume         Today
## Year   -0.030519101  0.84194162 -0.032459894
## Lag1   -0.008183096 -0.06495131 -0.075031842
## Lag2   -0.072499482 -0.08551314  0.059166717
## Lag3    0.060657175 -0.06928771 -0.071243639
## Lag4   -0.075675027 -0.06107462 -0.007825873
## Lag5    1.000000000 -0.05851741  0.011012698
## Volume -0.058517414  1.00000000 -0.033077783
## Today   0.011012698 -0.03307778  1.000000000
```

```r
plot(Weekly$Year,Weekly$Volume)
```



Answer:
We can see that variable `Year` and variable `Volume` has a strong linear relationship. While other variables' correlations among each other are all very small.

(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 appear to be statistically significant? If so, which ones?

```r
glm <- glm(Direction ~ Lag1 + Lag2 + Lag3 + Lag4 + Lag5 + Volume,
           data = Weekly, family = binomial)
summary(glm)
```

```
##
## 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
```

Answer:
The predictor `lag2` appears to be statistically significant with a p-value less than 0.05.

(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.

```r
glm.prob1 = predict(glm, type = 'response')
glm.predict1 = ifelse(glm.prob1 > 0.5, 'Up', 'Down')
table(glm.predict1, Weekly$Direction)
```

```
##
## glm.predict1 Down  Up
##         Down   54  48
##         Up    430 557
```

Answer:
The overall fraction of correct predictions is 56.1%. When actual direction goes down, model has a 88.8% error rate, which is the major mistake made by logistic regression; when actual direction goes up, model only has a 8.0% error rate.

(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).

```
training = Weekly %>% filter(Year<=2008)
test = Weekly %>% filter(Year>=2009)
glm2 <- glm(Direction ~ Lag2, data = training, family = binomial)
glm.prob2 = predict(glm2, test, type = 'response')
glm.predict2 = ifelse(glm.prob2 > 0.5, 'Up', 'Down')
table(glm.predict2, test$Direction)
```

```
##
## glm.predict2 Down Up
##         Down    9  5
##         Up     34 56
```

Answer:
The overall fraction of correct predictions is 62.5%.

(e) Repeat (d) using LDA.

```
lda=lda(Direction ~ Lag2, data=training)
lda.predict = predict(lda, test)
lda.class = lda.predict$class;
table(lda.class, test$Direction)
```

```
##
## lda.class Down Up
##      Down    9  5
##      Up     34 56
```

Answer:
The overall fraction of correct predictions is 62.5%.

(f) Repeat (d) using QDA.

```
qda=qda(Direction ~ Lag2, data=training)
qda.predict = predict(qda, test)
qda.class = qda.predict$class;
table(qda.class, test$Direction)
```

```
##
## qda.class Down Up
##      Down    0  0
##      Up     43 61
```

Answer:
The overall fraction of correct predictions is 58.7%.

(g) Repeat (d) using KNN with $K = 1$.

```
attach(Weekly)
train = (Year < 2009)
train.X = as.matrix(Lag2[train])
test.X = as.matrix(Lag2[!train])
train.Direction = Direction[train]
Direction.2009 = Direction[!train]
set.seed(1)
knn.pred = knn(train.X, test.X, train.Direction, k=1)
table(knn.pred, Direction.2009)
```

```
##          Direction.2009
## knn.pred Down Up
##     Down   21 30
##     Up     22 31
```

Answer:
The overall fraction of correct predictions is 50%.

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

Answer:
Logistic regression and LDA.

11.

In this problem, you will develop a model to predict whether a given car gets high or low gas mileage based on the Auto data set.

(a) Create a binary variable, `mpg01`, that contains a 1 if `mpg` contains a value above its median, and a 0 if `mpg` contains a value below its median. You can compute the median using the `median()` function. Note you may find it helpful to use the `data.frame()` function to create a single data set containing both `mpg01` and the other `Auto` variables.

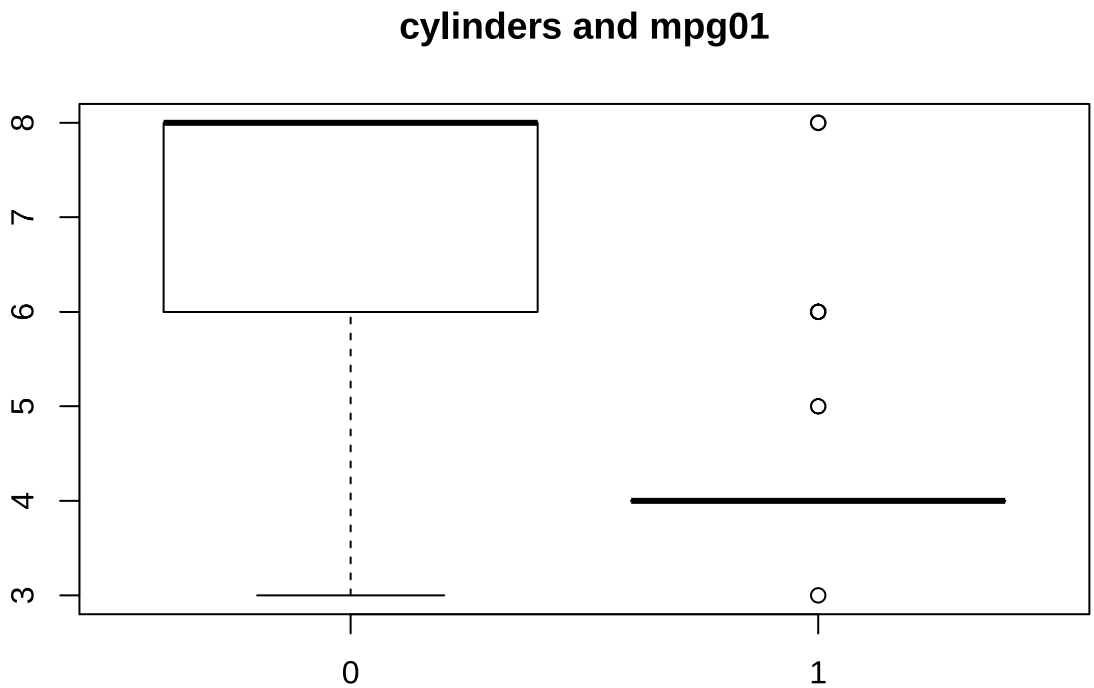
```r
attach(Auto)
mpg01 = rep(0, length(mpg))
mpg01[mpg > median(mpg)] = 1
Auto = data.frame(Auto, mpg01)
```

(b) Explore the data graphically in order to investigate the association between `mpg01` and the other features. Which of the other features seem most likely to be useful in predicting `mpg01`? Scatterplots and boxplots may be useful tools to answer this question. Describe your findings.

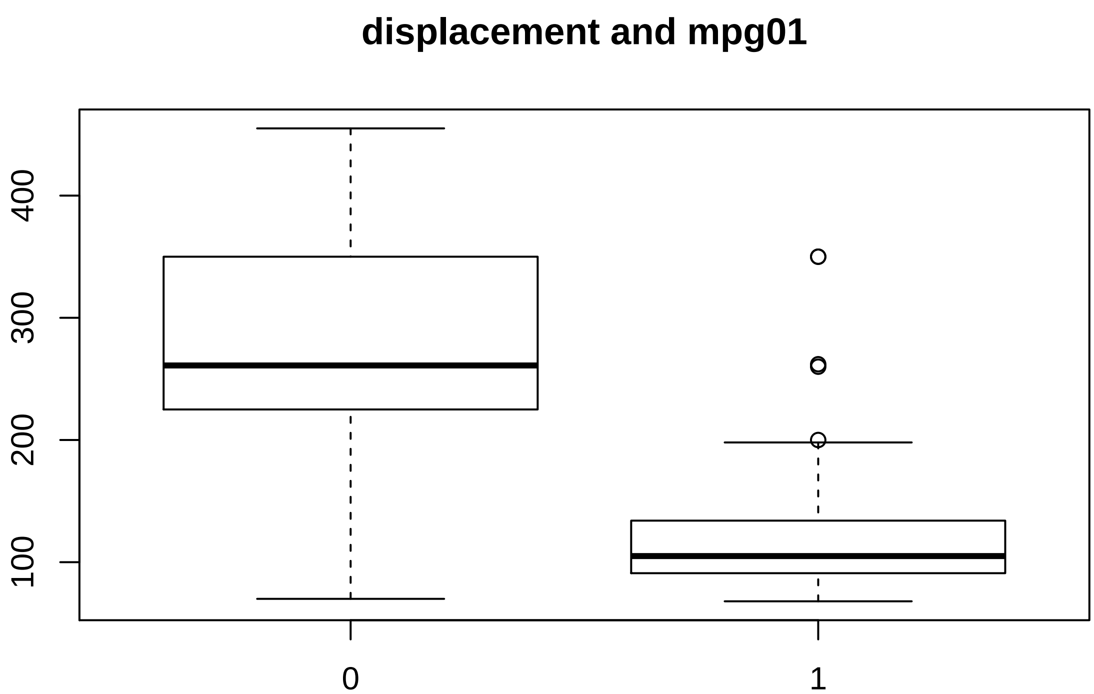
```r
cor(Auto[, -9])
```

```
##                     mpg  cylinders displacement horsepower     weight
## mpg           1.0000000 -0.7776175   -0.8051269 -0.7784268 -0.8322442
## cylinders    -0.7776175  1.0000000    0.9508233  0.8429834  0.8975273
## displacement -0.8051269  0.9508233    1.0000000  0.8972570  0.9329944
## horsepower   -0.7784268  0.8429834    0.8972570  1.0000000  0.8645377
## weight       -0.8322442  0.8975273    0.9329944  0.8645377  1.0000000
## acceleration  0.4233285 -0.5046834   -0.5438005 -0.6891955 -0.4168392
## year          0.5805410 -0.3456474   -0.3698552 -0.4163615 -0.3091199
## origin        0.5652088 -0.5689316   -0.6145351 -0.4551715 -0.5850054
## mpg01         0.8369392 -0.7591939   -0.7534766 -0.6670526 -0.7577566
##              acceleration       year     origin      mpg01
## mpg             0.4233285  0.5805410  0.5652088  0.8369392
## cylinders      -0.5046834 -0.3456474 -0.5689316 -0.7591939
## displacement   -0.5438005 -0.3698552 -0.6145351 -0.7534766
## horsepower     -0.6891955 -0.4163615 -0.4551715 -0.6670526
## weight         -0.4168392 -0.3091199 -0.5850054 -0.7577566
## acceleration    1.0000000  0.2903161  0.2127458  0.3468215
## year            0.2903161  1.0000000  0.1815277  0.4299042
## origin          0.2127458  0.1815277  1.0000000  0.5136984
## mpg01           0.3468215  0.4299042  0.5136984  1.0000000
```

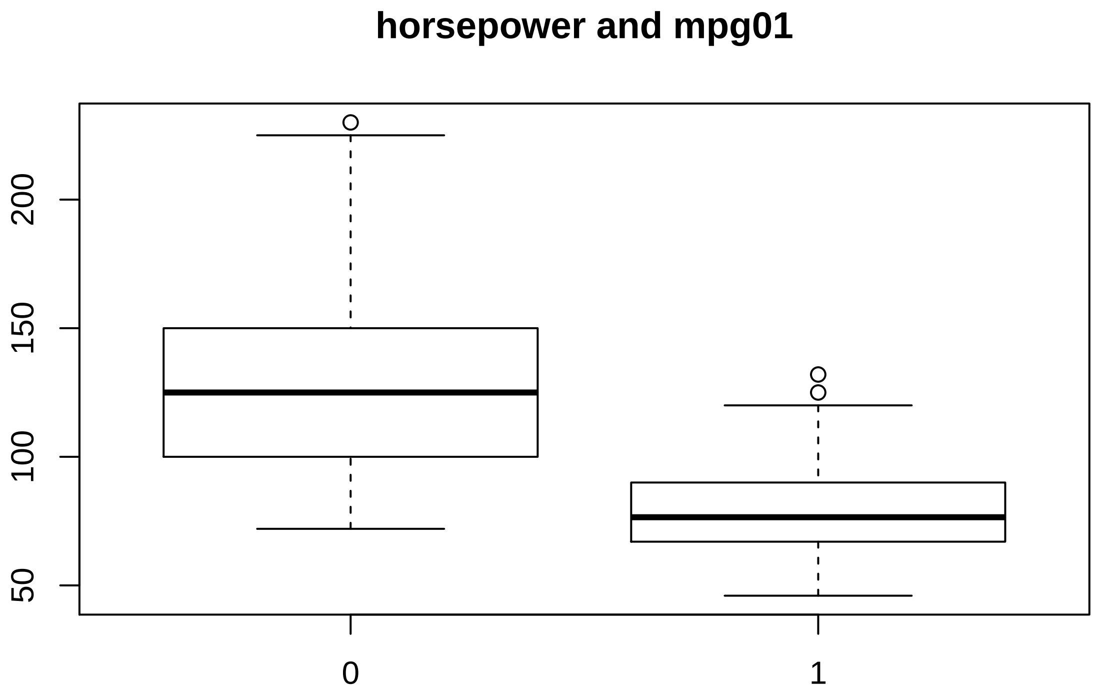
```r
boxplot(cylinders ~ mpg01, data = Auto, main = "cylinders and mpg01")
```



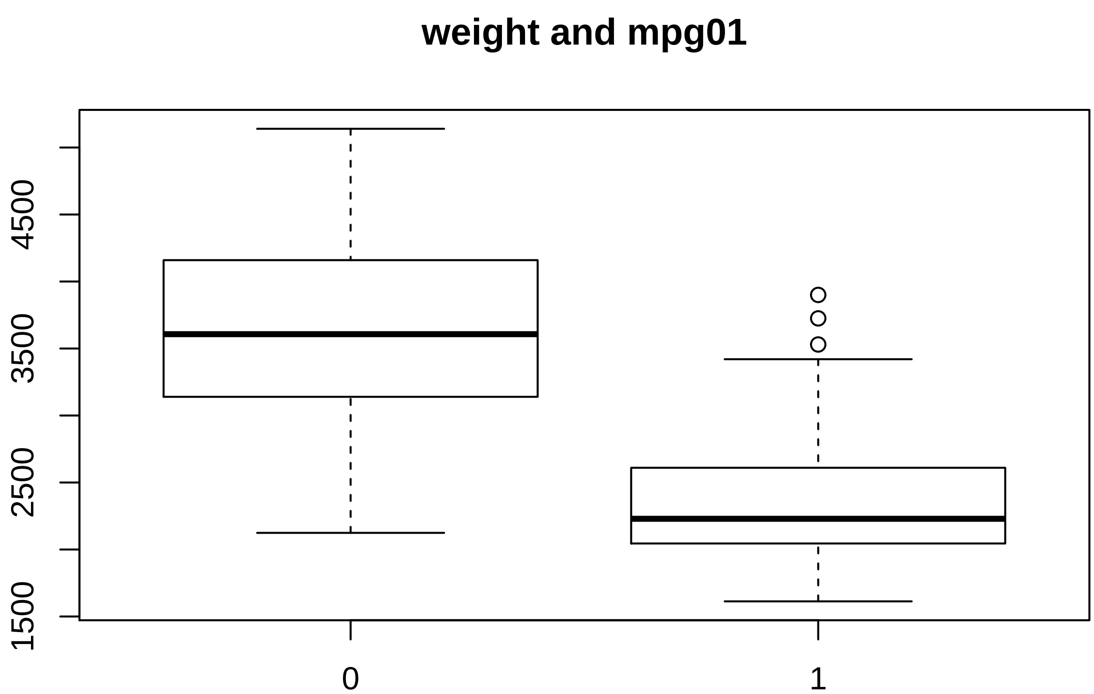

```r
boxplot(displacement ~ mpg01, data = Auto, main = "displacement and mpg01")
```




```r
boxplot(horsepower ~ mpg01, data = Auto, main = "horsepower and mpg01")
```




```r
boxplot(weight ~ mpg01, data = Auto, main = "weight and mpg01")
```



Answer:
`cylinders`, `displacement`, `horsepower`, `weight` could be useful in predicting `mpg01`.

(c) Split the data into a training set and a test set.

```r
training = Auto[1:200,]
test = Auto[201:392,]
```

(d) Perform LDA on the training data in order to predict `mpg01` using the variables that seemed most associated with `mpg01` in (b). What is the test error of the model obtained?

```r
lda=lda(mpg01 ~ cylinders + weight + displacement + horsepower, data=training)
lda.predict = predict(lda, test)
lda.class = lda.predict$class;
table(lda.class, test$mpg01)
```

```
##
## lda.class   0   1
##         0  56  12
##         1   8 116
```

Answer:
The test error of the model is 10.4%.

(e) Perform QDA on the training data in order to predict `mpg01` using the variables that seemed most associated with `mpg01` in (b). What is the test error of the model obtained?

```r
qda=qda(mpg01 ~ cylinders + weight + displacement + horsepower, data=training)
qda.predict = predict(qda, test)
qda.class = qda.predict$class;
table(qda.class, test$mpg01)
```

```
##
## qda.class   0   1
##         0  60  22
##         1   4 106
```

Answer:
The test error of the model is 13.5%.

(f) Perform logistic regression on the training data in order to predict `mpg01` using the variables that seemed most associated with `mpg01` in (b). What is the test error of the model obtained?

```r
glm <- glm(mpg01 ~ cylinders + weight + displacement + horsepower,
           data=training, family = binomial)
glm.prob = predict(glm, test, type = 'response')
glm.predict = ifelse(glm.prob > 0.5, '1', '0')
table(glm.predict, test$mpg01)
```

```
##
## glm.predict  0  1
##           0 61 36
##           1  3 92
```

Answer:
The test error of the model is 20.3%.

(g) Perform KNN on the training data, with several values of K, in order to predict mpg01. Use only the variables that seemed most associated with mpg01 in (b). What test errors do you obtain? Which value of K seems to perform the best on this data set?

```r
train = (rownames(Auto) <= 200)
test = !train
mpg01.test = mpg01[test]
train.X = cbind(cylinders, weight, displacement, horsepower)[train, ]
test.X = cbind(cylinders, weight, displacement, horsepower)[test, ]
train.mpg01 = mpg01[train]
set.seed(1)
for(i in 1:113){
```

```
  knn.pred = knn(train.X, test.X, train.mpg01, k = i)
  print(i)
  print(mean(knn.pred != mpg01.test))
}
```

```
## [1] 1
## [1] 0.1397849
## [1] 2
## [1] 0.1146953
## [1] 3
## [1] 0.1326165
## [1] 4
## [1] 0.125448
## [1] 5
## [1] 0.1218638
## [1] 6
## [1] 0.125448
## [1] 7
## [1] 0.1218638
## [1] 8
## [1] 0.125448
## [1] 9
## [1] 0.1146953
## [1] 10
## [1] 0.1182796
## [1] 11
## [1] 0.1146953
## [1] 12
## [1] 0.1146953
## [1] 13
## [1] 0.125448
## [1] 14
## [1] 0.1326165
## [1] 15
## [1] 0.1218638
## [1] 16
## [1] 0.125448
## [1] 17
## [1] 0.125448
## [1] 18
## [1] 0.125448
## [1] 19
## [1] 0.1218638
## [1] 20
## [1] 0.125448
## [1] 21
## [1] 0.125448
## [1] 22
## [1] 0.125448
## [1] 23
## [1] 0.125448
## [1] 24
## [1] 0.1326165
## [1] 25
```

```
## [1] 0.125448
## [1] 26
## [1] 0.125448
## [1] 27
## [1] 0.1182796
## [1] 28
## [1] 0.1182796
## [1] 29
## [1] 0.1182796
## [1] 30
## [1] 0.1290323
## [1] 31
## [1] 0.1326165
## [1] 32
## [1] 0.125448
## [1] 33
## [1] 0.1397849
## [1] 34
## [1] 0.1433692
## [1] 35
## [1] 0.1469534
## [1] 36
## [1] 0.1505376
## [1] 37
## [1] 0.1577061
## [1] 38
## [1] 0.1577061
## [1] 39
## [1] 0.1469534
## [1] 40
## [1] 0.1469534
## [1] 41
## [1] 0.1469534
## [1] 42
## [1] 0.1469534
## [1] 43
## [1] 0.1433692
## [1] 44
## [1] 0.1469534
## [1] 45
## [1] 0.1397849
## [1] 46
## [1] 0.1469534
## [1] 47
## [1] 0.1362007
## [1] 48
## [1] 0.1362007
## [1] 49
## [1] 0.1397849
## [1] 50
## [1] 0.1433692
## [1] 51
## [1] 0.1433692
## [1] 52
```

```
## [1] 0.1433692
## [1] 53
## [1] 0.1433692
## [1] 54
## [1] 0.1397849
## [1] 55
## [1] 0.1362007
## [1] 56
## [1] 0.1326165
## [1] 57
## [1] 0.125448
## [1] 58
## [1] 0.1290323
## [1] 59
## [1] 0.1218638
## [1] 60
## [1] 0.1218638
## [1] 61
## [1] 0.1290323
## [1] 62
## [1] 0.1290323
## [1] 63
## [1] 0.1326165
## [1] 64
## [1] 0.1362007
## [1] 65
## [1] 0.1326165
## [1] 66
## [1] 0.1326165
## [1] 67
## [1] 0.1290323
## [1] 68
## [1] 0.1290323
## [1] 69
## [1] 0.1218638
## [1] 70
## [1] 0.1182796
## [1] 71
## [1] 0.1146953
## [1] 72
## [1] 0.1146953
## [1] 73
## [1] 0.1146953
## [1] 74
## [1] 0.1111111
## [1] 75
## [1] 0.1146953
## [1] 76
## [1] 0.1146953
## [1] 77
## [1] 0.1146953
## [1] 78
## [1] 0.1146953
## [1] 79
```

```
## [1] 0.1146953
## [1] 80
## [1] 0.1218638
## [1] 81
## [1] 0.1146953
## [1] 82
## [1] 0.1146953
## [1] 83
## [1] 0.1146953
## [1] 84
## [1] 0.1146953
## [1] 85
## [1] 0.1146953
## [1] 86
## [1] 0.1111111
## [1] 87
## [1] 0.1111111
## [1] 88
## [1] 0.1218638
## [1] 89
## [1] 0.1433692
## [1] 90
## [1] 0.3189964
## [1] 91
## [1] 0.5412186
## [1] 92
## [1] 0.5412186
## [1] 93
## [1] 0.5412186
## [1] 94
## [1] 0.5412186
## [1] 95
## [1] 0.5412186
## [1] 96
## [1] 0.5412186
## [1] 97
## [1] 0.5412186
## [1] 98
## [1] 0.5412186
## [1] 99
## [1] 0.5412186
## [1] 100
## [1] 0.5412186
## [1] 101
## [1] 0.5412186
## [1] 102
## [1] 0.5412186
## [1] 103
## [1] 0.5412186
## [1] 104
## [1] 0.5412186
## [1] 105
## [1] 0.5412186
## [1] 106
```

```
## [1] 0.5412186
## [1] 107
## [1] 0.5412186
## [1] 108
## [1] 0.5412186
## [1] 109
## [1] 0.5412186
## [1] 110
## [1] 0.5412186
## [1] 111
## [1] 0.5412186
## [1] 112
## [1] 0.5412186
## [1] 113
## [1] 0.5412186
```

Answer:
When $K = 75, 87, 88$, it seems to perform the best on this data set.