FE590. Assignment #2

Gang Ping Zhu

2017-10-14

Instructions

In this assignment, you should use R markdown to answer the questions below. Simply type your R code into embedded chunks as shown above. When you have completed the assignment, knit the document into a PDF file, and upload both the .pdf and .Rmd files to Canvas.

Question 1 (based on JWHT Chapter 2, Problem 9)

Use the Auto data set from the textbook's website. When reading the data, use the options as.is = TRUE and na.strings="?". Remove the unavailable data using the na.omit() function.

```
setwd("C:/Users/gang.ping.m.zhu/OneDrive - Accenture/Stevens/FE 590")
auto <- read.csv("Auto.csv", as.is = TRUE, na.strings = "?")
auto <- na.omit(auto)
head(auto)
```

```
##   mpg cylinders displacement horsepower weight acceleration year origin
## 1  18         8          307        130   3504         12.0   70      1
## 2  15         8          350        165   3693         11.5   70      1
## 3  18         8          318        150   3436         11.0   70      1
## 4  16         8          304        150   3433         12.0   70      1
## 5  17         8          302        140   3449         10.5   70      1
## 6  15         8          429        198   4341         10.0   70      1
##                        name
## 1 chevrolet chevelle malibu
## 2         buick skylark 320
## 3        plymouth satellite
## 4             amc rebel sst
## 5               ford torino
## 6          ford galaxie 500
```

1. List the names of the variables in the data set.

```
colnames(auto)
```

```
## [1] "mpg"          "cylinders"    "displacement" "horsepower"
## [5] "weight"       "acceleration" "year"         "origin"
## [9] "name"
```

2. The columns origin and name are unimportant variables. Create a new data frame called cars that contains none of these unimportant variables

```
cars <- subset(auto, select = c(1,2,3,4,5,6,7))
head(cars)
```

```
##   mpg cylinders displacement horsepower weight acceleration year
## 1  18         8          307        130   3504         12.0   70
## 2  15         8          350        165   3693         11.5   70
## 3  18         8          318        150   3436         11.0   70
## 4  16         8          304        150   3433         12.0   70
## 5  17         8          302        140   3449         10.5   70
## 6  15         8          429        198   4341         10.0   70
```

3. What is the range of each quantitative variable? Answer this question using the range() function with the sapply() function (e.g., sapply(cars, range). Print a simple table of the ranges of the variables. The rows should correspond to the variables. The first column should be the lowest value of the corresponding variable, and the second column should be the maximum value of the variable. The columns should be suitably labeled.

```
rangecars <- sapply(cars, range)
rangecars <- as.data.frame(rangecars)
rangecars
```

```
##    mpg cylinders displacement horsepower weight acceleration year
## 1  9.0         3           68         46   1613          8.0   70
## 2 46.6         8          455        230   5140         24.8   82
```

4. What is the mean and standard deviation of each variable? Create a simple table of the means and standard deviations.

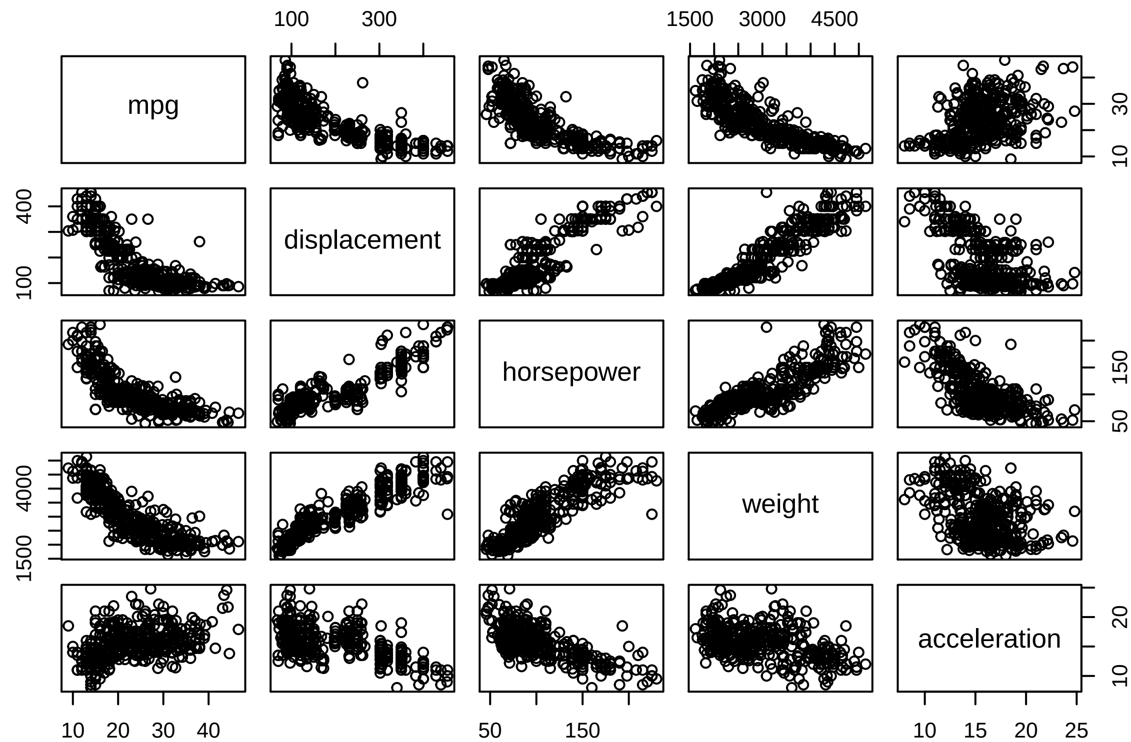
```
meancars <- sapply(cars, mean)
sdcars <- sapply(cars, sd)

msd <- rbind(meancars, sdcars)
msd
```

```
##                mpg cylinders displacement horsepower    weight
## meancars 23.445918  5.471939      194.412  104.46939 2977.5842
## sdcars    7.805007  1.705783      104.644   38.49116  849.4026
##          acceleration      year
## meancars    15.541327 75.979592
## sdcars       2.758864  3.683737
```

5. Create a scatterplot matrix that includes the variables mpg, displacement, horsepower, weight, and acceleration using the pairs() function.

```r
pairs(~ mpg + displacement + horsepower + weight + acceleration, data = cars)
```



6. From the scatterplot, it should be clear that mpg has an almost linear relationship to predictors, and higher-order relationships to other variables. Using the regsubsets function in the leaps library, regress mpg onto

- displacement
- displacement squared
- horsepower
- horsepower squared
- weight
- weight squared
- acceleration

```r
library("leaps")
cars$displacement.squared <- cars$displacement^2
cars$horsepower.squared <- cars$horsepower^2
cars$weight.squared <- cars$weight^2
```

```r
a <- regsubsets(mpg~., data=cars)
a
```

```
## Subset selection object
## Call: regsubsets.formula(mpg ~ ., data = cars)
## 9 Variables  (and intercept)
##                      Forced in Forced out
## cylinders                FALSE      FALSE
## displacement             FALSE      FALSE
## horsepower               FALSE      FALSE
## weight                   FALSE      FALSE
## acceleration             FALSE      FALSE
## year                     FALSE      FALSE
## displacement.squared     FALSE      FALSE
## horsepower.squared       FALSE      FALSE
## weight.squared           FALSE      FALSE
## 1 subsets of each size up to 8
## Selection Algorithm: exhaustive
```

Print a table showing what variables would be selected using best subset selection for all model orders.

```r
summary(a)
```

```
## Subset selection object
## Call: regsubsets.formula(mpg ~ ., data = cars)
## 9 Variables  (and intercept)
##                      Forced in Forced out
## cylinders                FALSE      FALSE
## displacement             FALSE      FALSE
## horsepower               FALSE      FALSE
## weight                   FALSE      FALSE
## acceleration             FALSE      FALSE
## year                     FALSE      FALSE
## displacement.squared     FALSE      FALSE
## horsepower.squared       FALSE      FALSE
## weight.squared           FALSE      FALSE
## 1 subsets of each size up to 8
## Selection Algorithm: exhaustive
##          cylinders displacement horsepower weight acceleration year
## 1  ( 1 ) " "       " "          " "        "*"    " "          " " 
## 2  ( 1 ) " "       " "          " "        "*"    " "          "*" 
## 3  ( 1 ) " "       " "          " "        "*"    " "          "*" 
## 4  ( 1 ) " "       " "          "*"        "*"    " "          "*" 
## 5  ( 1 ) " "       " "          "*"        "*"    " "          "*" 
## 6  ( 1 ) " "       " "          "*"        "*"    "*"          "*" 
## 7  ( 1 ) " "       "*"          "*"        "*"    " "          "*" 
## 8  ( 1 ) "*"       "*"          "*"        "*"    " "          "*" 
##          displacement.squared horsepower.squared weight.squared
## 1  ( 1 ) " "                  " "                " "           
## 2  ( 1 ) " "                  " "                " "           
## 3  ( 1 ) " "                  " "                "*"           
## 4  ( 1 ) " "                  " "                "*"           
## 5  ( 1 ) " "                  "*"                "*"           
## 6  ( 1 ) " "                  "*"                "*"           
## 7  ( 1 ) "*"                  "*"                "*"           
```

```
## 8  ( 1 ) "*"                    "*"                  "*"
```

```r
t(summary(a)$which)
```

```
##                          1     2     3     4     5     6     7     8
## (Intercept)           TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
## cylinders            FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE
## displacement         FALSE FALSE FALSE FALSE FALSE FALSE  TRUE  TRUE
## horsepower           FALSE FALSE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE
## weight                TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
## acceleration         FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE
## year                 FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
## displacement.squared FALSE FALSE FALSE FALSE FALSE FALSE  TRUE  TRUE
## horsepower.squared   FALSE FALSE FALSE FALSE  TRUE  TRUE  TRUE  TRUE
## weight.squared       FALSE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
```

What is the most important variable affecting fuel consumption?

```r
# weight
```

What is the second most important variable affecting fuel consumption?

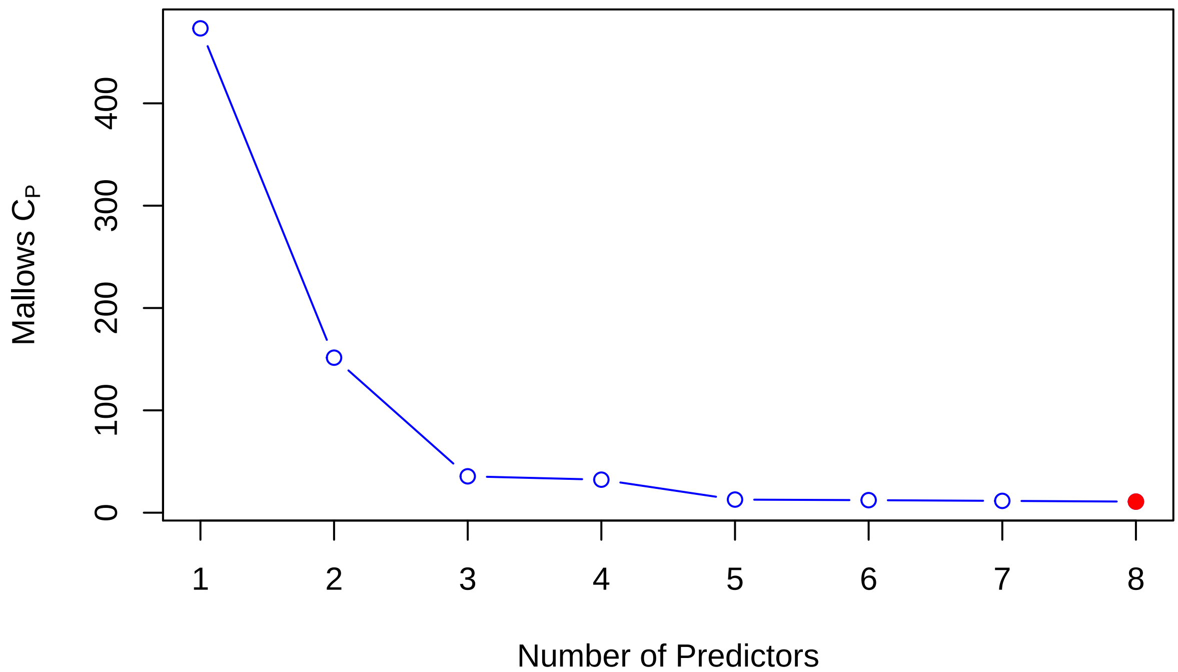
```r
# year
```

What is the third most important variable affecting fuel consumption?

```r
# horsepower
```

7. Plot a graph showing Mallow's Cp as a function of the order of the model. Which model is the best?

```r
cp=summary(a)$cp
i=which.min(cp)
plot(cp,type='b',col="blue",xlab="Number of Predictors",ylab=expression("Mallows C"[P]))
points(i,cp[i],pch=19,col="red")
```



Based on the chart above, we can see that our last model is best out of the subset of predictors. The small value of C[P] means that the model is relatively precise.

Question 2 (based on JWHT Chapter 3, Problem 10)

This exercise involves the Boston housing data set.

1. Load in the Boston data set, which is part of the MASS library in R. The data set is contained in the object Boston. Read about the data set using the command ?Boston. How many rows are in this data set? How many columns? What do the rows and columns represent?

```
library("MASS")
names(Boston)
```

```
##  [1] "crim"    "zn"      "indus"   "chas"    "nox"     "rm"      "age"
##  [8] "dis"     "rad"     "tax"     "ptratio" "black"   "lstat"   "medv"
```

```
?Boston
```

The Boston data frame has 506 rows and 14 columns.The rows represent housing values in the suburbs of Boston. The columns represent different attributes of the suburbs of Boston. They are represented by the following:

- crim - per capita crime rate by town.

- zn - proportion of residential land zoned for lots over 25,000 sq.ft.
- indus - proportion of non-retail business acres per town.
- chas - Charles River dummy variable (= 1 if tract bounds river; 0 otherwise).
- nox - nitrogen oxides concentration (parts per 10 million).
- rm - average number of rooms per dwelling.
- age - proportion of owner-occupied units built prior to 1940.
- dis - weighted mean of distances to five Boston employment centres.
- rad - index of accessibility to radial highways.
- tax - full-value property-tax rate per $10,000.
- ptratio - pupil-teacher ratio by town.
- black - 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town.
- lstat - lower status of the population (percent).
- medv - median value of owner-occupied homes in $1000s.

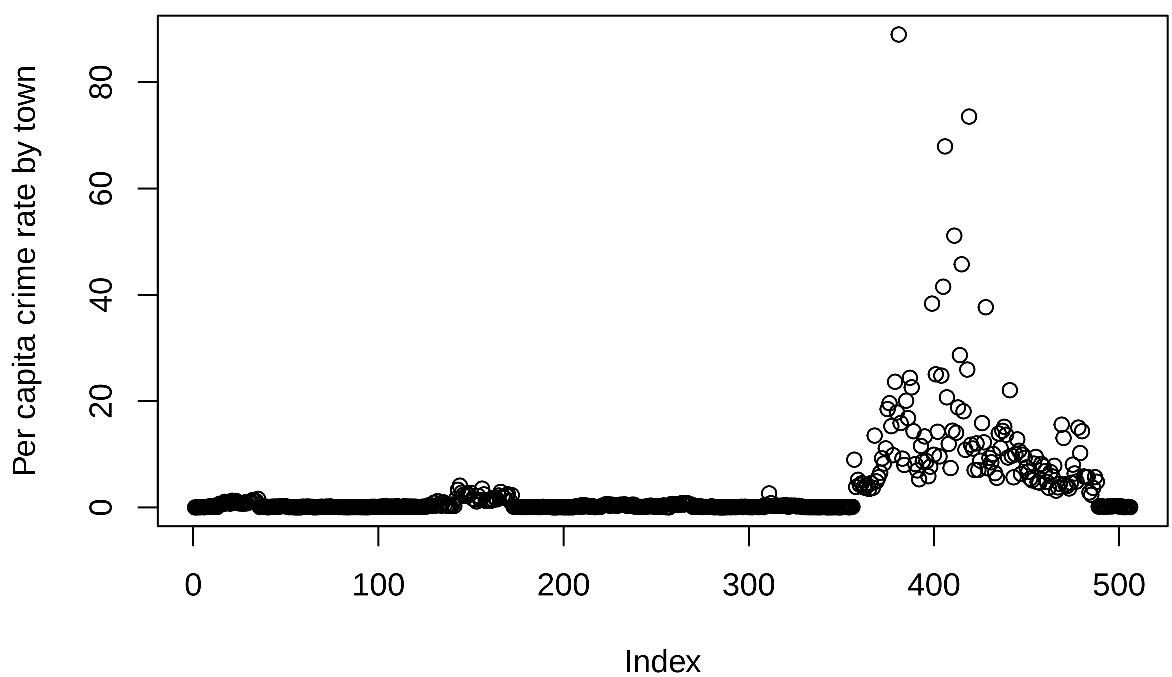
2. Do any of the suburbs of Boston appear to have particularly high crime rates?

Based on the chart below, there are a few areas where crime seems to be particularly high.

```r
attach(Boston)
summary(Boston$crim)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##  0.00632  0.08204  0.25651  3.61352  3.67708 88.97620
```

```r
plot(Boston$crim, ylab = "Per capita crime rate by town")
```

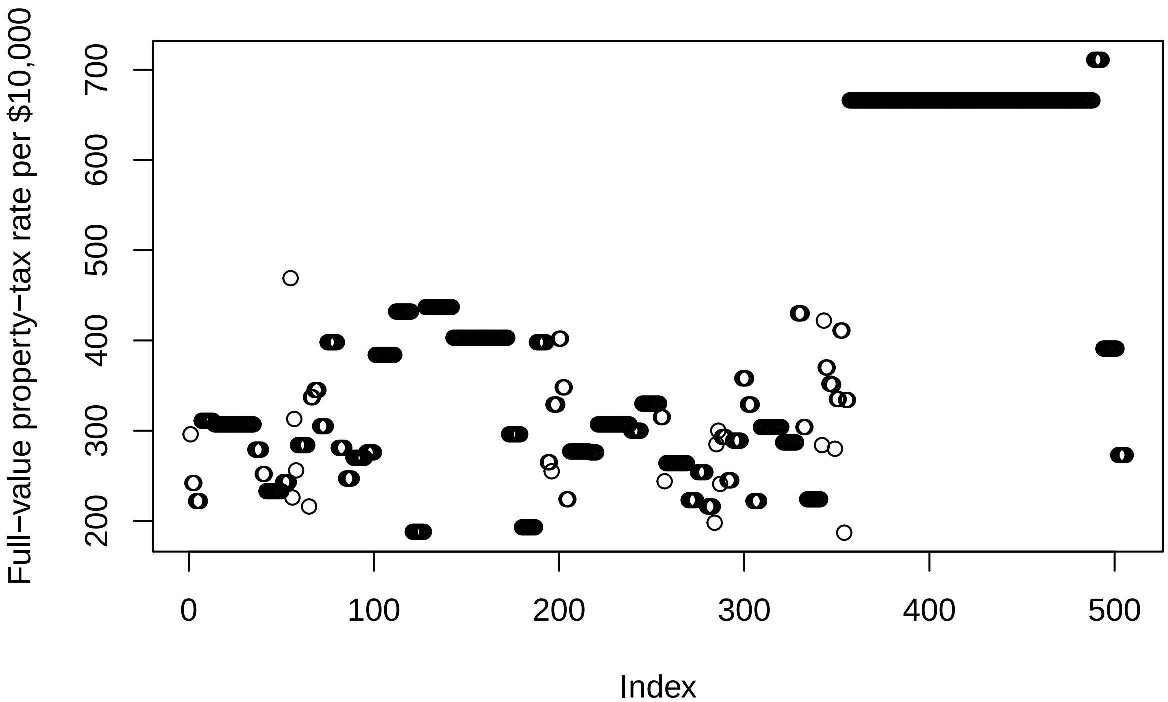


Tax rates? Based on the chart below, we can see there are a few areas where tax seems to be particularly high. It’s placed on same areas where the crime rate is high.

```r
summary(Boston$tax)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   187.0   279.0   330.0   408.2   666.0   711.0
```

```r
plot(Boston$tax, ylab = "Full-value property-tax rate per $10,000")
```

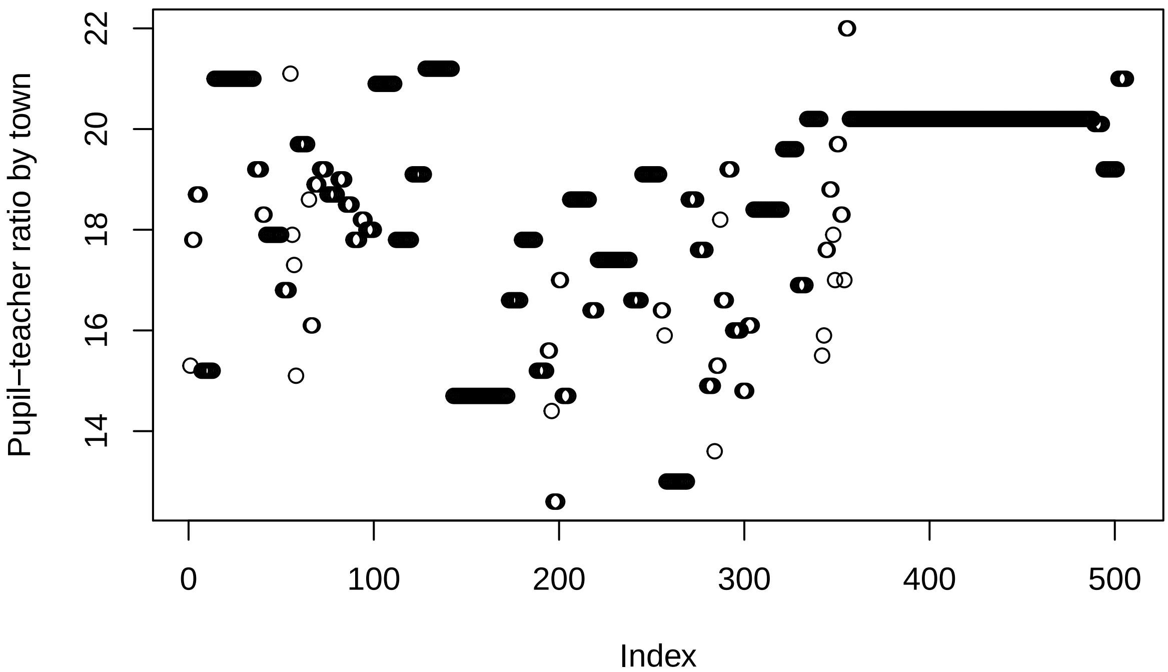


Pupil-teacher ratios? Based on the chart below, we can see there are a few areas where pupil to teacher ratio seems to be high. Unlike the high crime rate and high tax rate, the same places that have high crime and high tax also seem to have a high pupil to teacher ratio but unlike those two attributes, there also seem to be other areas that have a high pupil to teacher ratio that don’t exhibit high tax or high crime.

```r
summary(Boston$ptratio)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   12.60   17.40   19.05   18.46   20.20   22.00
```

```r
plot(Boston$ptratio, ylab = "Pupil-teacher ratio by town")
```



Comment on the range of each predictor. While there is a range for the crime rate, most of the values for crime are pretty low with a portion of it being disparate from the rest. The tax rate seems to follow a similar pattern but it shows a bit more diversity with a spike in tax in the area where crime is high. As for the puil-teacher ratio, that range is very diverse even though we see a spike in the same area as the high tax and crime rate

3. How many of the suburbs in this data set bound the Charles river?

```
table(Boston$chas)
```

```
##
##   0   1
## 471  35
```

There are 35 suburbs where tract bounds the Charles river.

4. What is the median pupil-teacher ratio among the towns in this data set?

```
median(Boston$ptratio)
```

```
## [1] 19.05
```

5. In this data set, how many of the suburbs average more than seven rooms per dwelling?

```r
table(Boston$rm > 7)
```

```
##
## FALSE  TRUE
##   442    64
```

There are 64 suburbs that average more than seven rooms per dwelling

More than eight rooms per dwelling?

```r
table(Boston$rm > 8)
```

```
##
## FALSE  TRUE
##   493    13
```

There are 13 suburbs that average than eight rooms per dwelling.

Comment on the suburbs that average more than eight rooms per dwelling.

```r
rms8 <- subset.data.frame(Boston, rm > 8)
summary(rms8)
```

```
##       crim                zn            indus             chas       
##  Min.   :0.02009   Min.   : 0.00   Min.   : 2.680   Min.   :0.0000  
##  1st Qu.:0.33147   1st Qu.: 0.00   1st Qu.: 3.970   1st Qu.:0.0000  
##  Median :0.52014   Median : 0.00   Median : 6.200   Median :0.0000  
##  Mean   :0.71879   Mean   :13.62   Mean   : 7.078   Mean   :0.1538  
##  3rd Qu.:0.57834   3rd Qu.:20.00   3rd Qu.: 6.200   3rd Qu.:0.0000  
##  Max.   :3.47428   Max.   :95.00   Max.   :19.580   Max.   :1.0000  
##       nox               rm             age             dis       
##  Min.   :0.4161   Min.   :8.034   Min.   : 8.40   Min.   :1.801  
##  1st Qu.:0.5040   1st Qu.:8.247   1st Qu.:70.40   1st Qu.:2.288  
##  Median :0.5070   Median :8.297   Median :78.30   Median :2.894  
##  Mean   :0.5392   Mean   :8.349   Mean   :71.54   Mean   :3.430  
##  3rd Qu.:0.6050   3rd Qu.:8.398   3rd Qu.:86.50   3rd Qu.:3.652  
##  Max.   :0.7180   Max.   :8.780   Max.   :93.90   Max.   :8.907  
##       rad              tax           ptratio          black      
##  Min.   : 2.000   Min.   :224.0   Min.   :13.00   Min.   :354.6  
##  1st Qu.: 5.000   1st Qu.:264.0   1st Qu.:14.70   1st Qu.:384.5  
##  Median : 7.000   Median :307.0   Median :17.40   Median :386.9  
##  Mean   : 7.462   Mean   :325.1   Mean   :16.36   Mean   :385.2  
##  3rd Qu.: 8.000   3rd Qu.:307.0   3rd Qu.:17.40   3rd Qu.:389.7  
##  Max.   :24.000   Max.   :666.0   Max.   :20.20   Max.   :396.9  
##      lstat           medv     
##  Min.   :2.47   Min.   :21.9  
##  1st Qu.:3.32   1st Qu.:41.7  
##  Median :4.14   Median :48.3  
##  Mean   :4.31   Mean   :44.2  
##  3rd Qu.:5.12   3rd Qu.:50.0  
##  Max.   :7.44   Max.   :50.0
```

These suburbs don't have much crime and are not taxed at the highest level. There are a good percentage of homes in these suburbs that are built prior to 1940.

Question 3 (based on JWHT Chapter 4, Problem 10)

This question should be answered using the Weekly data set, which is part of the ISLR package. This data contains 1,089 weekly returns for 21 years, from the beginning of 1990 to the end of 2010.

1. What does the data represent?

```
library("ISLR")
attach(Weekly)
?Weekly
```

This Weekly data represents the weekly percentage returns for the S&P 500 stock index between 1990 and 2010.

2. 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?

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

```
##
## 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
```

Yes, there appears to be a significant Coefficient. Lag2 appears to be statistically significant.

3. Fit a 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).

```
library(class)
trainweekly <- subset(Weekly, Year < 2009)
otherweekly <- subset(Weekly, Year > 2008)
glm.fit=glm(Direction~Lag2,family=binomial,data=trainweekly)
summary(glm.fit)
```

```
##
## Call:
## glm(formula = Direction ~ Lag2, family = binomial, data = trainweekly)
##
## 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
```

```
glm.fit=glm(Direction~Lag2,family=binomial,data=otherweekly)
glm.probs=predict(glm.fit,type="response")
glm.pred=rep("Down",104)
glm.pred[glm.probs>.5]="Up"
table(glm.pred,otherweekly$Direction)
```

```
##
## glm.pred Down Up
##     Down    8  4
##     Up     35 57
```

4. Repeat Part 3 using LDA.

```
lda.fit=lda(Direction~Lag2,data=trainweekly)
summary(lda.fit)
```

```
##         Length Class  Mode
## prior   2      -none- numeric
## counts  2      -none- numeric
## means   2      -none- numeric
## scaling 1      -none- numeric
## lev     2      -none- character
## svd     1      -none- numeric
## N       1      -none- numeric
## call    3      -none- call
## terms   3      terms  call
## xlevels 0      -none- list
```

```
lda.fit <- lda(Direction~Lag2,data=otherweekly)
lda.pred <- predict(lda.fit,otherweekly)
lda.class <- lda.pred$class
table(lda.class,otherweekly$Direction)
```

```
##
## lda.class Down Up
##      Down    8  4
##      Up     35 57
```

5. Repeat Part 3 using QDA.

```
qda.fit <- qda(Direction~Lag2,data=trainweekly)
summary(qda.fit)
```

```
##         Length Class  Mode
## prior   2      -none- numeric
## counts  2      -none- numeric
## means   2      -none- numeric
## scaling 2      -none- numeric
## ldet    2      -none- numeric
## lev     2      -none- character
## N       1      -none- numeric
## call    3      -none- call
## terms   3      terms  call
## xlevels 0      -none- list
```

```
qda.fit <- qda(Direction~Lag2,data=otherweekly)
qda.class <- predict(qda.fit,otherweekly)$class
table(qda.class,otherweekly$Direction)
```

```
##
## qda.class Down Up
##      Down    9  4
##      Up     34 57
```

6. Repeat Part 3 using KNN with K = 1, 2, 3. (Fit a 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))

```
train.X <- as.data.frame(trainweekly$Lag2)
test.X <- as.data.frame(otherweekly$Lag2)
train.Direction=trainweekly$Direction

set.seed(1)
knn.pred=knn(train.X,test.X,train.Direction, k=1)
table(knn.pred,otherweekly$Direction)
```

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

```
knn.pred=knn(train.X,test.X,train.Direction, k=2)
table(knn.pred,otherweekly$Direction)
```

```
##
## knn.pred Down Up
##     Down   18 25
##     Up     25 36
```

```
knn.pred=knn(train.X,test.X,train.Direction, k=3)
table(knn.pred,otherweekly$Direction)
```

```
##
## knn.pred Down Up
##     Down   16 20
##     Up     27 41
```

7. Which of these methods in Parts 3, 4, 5, and 6 appears to provide the best results on this data?

```
#QDA appears to have the best results out of the different methods with a 63% accuracy.
```

Question 4

Write a function that works in R to gives you the parameters from a linear regression on a data set between two sets of values (in other words you only have to do the 2-D case). Include in the output the standard error of your variables. You cannot use the lm command in this function or any of the other built in regression models. For example your output could be a 2x2 matrix with the parameters in the first column and the standard errors in the second column. For up to 5 bonus points, format your output so that it displays and operates similar in function to the output of the lm command.(i.e. in a data frame that includes all potentially useful outputs)

```
#y=X??+??
#?????N(0,(??^2)I)
lnreg <- function(x, y){
  x1 <- as.matrix(x)
  y1 <- cbind(constant = 1, as.matrix(y))
  vb <- solve(t(y1)%*%y1, t(y1)%*%x1)
  ds <- sum((x1 - y1%*%vb)^2)/(nrow(y1)-ncol(x1))
  StdErrors <- sqrt(diag(ds*chol2inv(chol(t(y1)%*%y1))))
  res <- cbind(vb, StdErrors)
  print(res)
}
```

Compare the output of your function to that of the lm command in R.

```
lnreg(Lag1, Lag2)
```

```
##                        StdErrors
## constant  0.16189250 0.07137051
##          -0.07484538 0.03022880
```

```
testlm <- lm(Lag1 ~ Lag2, data = Weekly)
testlm
```

```
##
## Call:
## lm(formula = Lag1 ~ Lag2, data = Weekly)
##
## Coefficients:
## (Intercept)         Lag2
##     0.16189     -0.07485
```

```
summary(testlm)
```

```
##
## Call:
## lm(formula = Lag1 ~ Lag2, data = Weekly)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -19.0604  -1.2715   0.1134   1.2796  11.2362
```

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.16189    0.07140   2.267   0.0236 *
## Lag2        -0.07485    0.03024  -2.475   0.0135 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.351 on 1087 degrees of freedom
## Multiple R-squared:  0.005603,   Adjusted R-squared:  0.004688
## F-statistic: 6.125 on 1 and 1087 DF,  p-value: 0.01348
```