# COMS 573 Homework 1

Boudhayan Banerjee

March 9, 2016

1. a. Euclidean distance between a given test point $(p_1, p_2, p_3)$ and the observation point $(q_1, q_2, q_3)$ is $d(p,q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + (p_3 - q_3)^2}$.

   Applying this formula to the given observation points we get,

   1. 3
   2. $\sqrt{5} = 2.236$
   3. $\sqrt{10} = 3.162$
   4. $\sqrt{5} = 2.236$
   5. $\sqrt{2} = 1.414$
   6. $\sqrt{3} = 1.732$

   b. If $k = 1$ then the smallest distance between the observation and test point is 1.414 i.e the $5th$ observation.Therefore prediction is **Green**.

   c. If $k = 3$ then the 3 nearest points from the test point (0,0,0) is observation point 2,4,5 and 6. Now point 2 and 4 have equal distance from the test point. But if we consider the point with smaller index then we will choose point 2,4 and 5.Now as 2 and 6 are Red therefore the prediction is **Red**.

   d. If Bayes decision boundary in this problem is highly non linear then value of k will be small. Because when k is small the decision boundary is overly flexible and find patterns in the data that don't correspond to the Bayes decision boundary.

   e. Following is the program which performs k-nearest neighbour classification for the given problem:
   # training points where response is Red:
   $R1 = c(2,0,1)$
   $R2 = c(0,1,3)$

```
R3 = c(1,−1,1)

# training points where response is Green:
G1 = c(0,3,0)
G2 = c(0,−1,2)
G3 = c(−1,0,1)

# train the cases
train=rbind(R1,R2,R3, G1,G2,G3)

#create the vector for classification label
cl=factor(c(rep("Red",3),c(rep("Green",3)))

#test the object to be classified
test=c(0,0,0)

#load class package that uses knn() function
library(class)

#call knn function and get summary
summary(knn(train,test,cl,k=3)
```

**output:**

```
Red Green
0 1
```

2. We have,

$$Y_i = \beta_1 x_i + e_i$$

It is given that $E[e_i] = 0, Var[e_i] = \sigma_e^2$ and $Cov[e_i, e_j] = 0 \ \forall i \neq j$
We need to find the least square estimator $\hat{\beta}_1$ of slope $\beta_1$.
Let $\hat{Y}_i = \beta_1 x_i$ be the estimator of Y on the ith value of X.
Now, $e_i = Y_i - \hat{Y}_i$
This is the difference between the observed value and the estimated value by our linear model for ith observation.The sum of the residual squares=$e_1^2 + e_2^2 + e_3^2 + e_4^2 + ....$
$= (Y_1 - \beta_1 x_1)^2 + (Y_2 - \beta_1 x_2)^2 + (Y_3 - \beta_1 x_3)^2 + ... + (Y_n - \beta_1 x_n)^2$
$= \Sigma_{i=1}^n (Y_i - \beta_1 x_i)^2$

We can have the least square estimator $\hat{\beta}_1$ by differentiating the above equation w.r.t $\beta_1$ and equating it to 0.
$\frac{\partial}{\partial \beta_1}(\Sigma_{i=1}^n (Y_i - \beta_1 x_i)^2) = 0$

$\rightarrow -2(\Sigma_{i=1}^{n}(Y_i - \beta_1 x_i))x_i = 0$
$\rightarrow \Sigma_{i=1}^{n} Y_i x_i = \Sigma_{i=1}^{n} \hat{\beta}_1 x_i$
$\rightarrow \Sigma_{i=1}^{n} Y_i x_i / \Sigma_{i=1}^{n} x_i^2 = \hat{\beta}$

$\hat{\beta}_1 = \Sigma_{i=1}^{n} Y_i x_i / \Sigma_{i=1}^{n} x_i^2.$

3. a. $> library(MASS)$
$>?Boston$

The Boston data set has 506 rows and 14 columns.
The rows represent the Housing values in Boston and columns represents the attributes of each House.

b. $pairs(Boston)$

[The plot is in page 8]

We have the following findings from the plots:
⇒ crim has correlation with age,dis,rad,tax and pratio
⇒ zn has correlation with indus,nox,age,lstat
⇒ indus has correlation with age and dis
⇒ nox has correlation with age and dis
⇒ dis has correlation with lstat
⇒ lstat has correlation with medv

c. To find the association between per capita crime rate and other attributes we need to do pairwise plotting.
Among 13 attributes only 5 has association with per capita crime rate.

$plot(Boston\$age, Boston\$crim)$
**If the age of the housings are high crime rate is also high**

$plot(Boston\$dis, Boston\$crim)$
**If the district is closer to the work locations then crime rate is high**

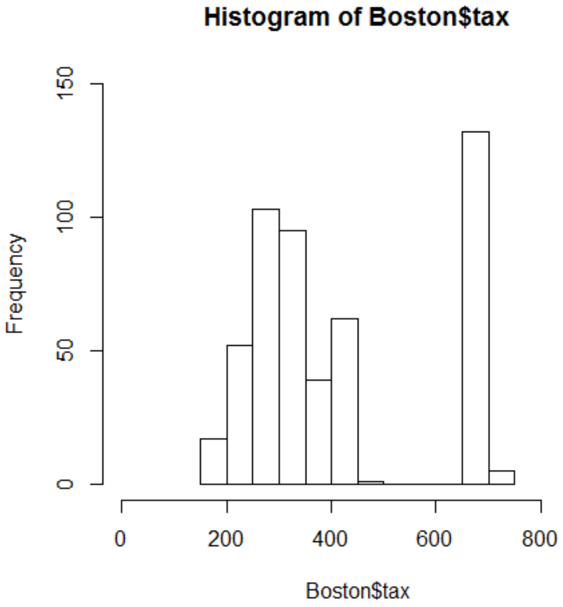
$plot(Boston\$rad, Boston\$crim)$
**Higher index of accessibility to radial highways means more crime**

$plot(Boston\$tax, Boston\$crim)$
**If the tax rate is higher the crim rate is also higher**

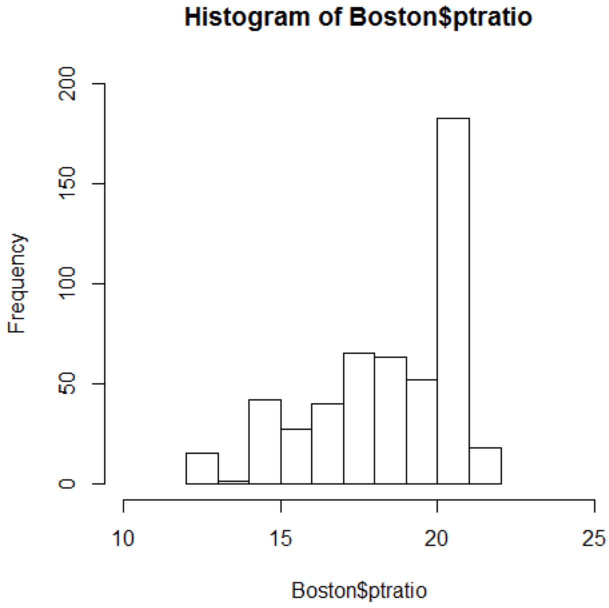
$plot(Boston\$ptratio, Boston\$crim)$
**If the pupil teacher ratio is higher crime rate is also higher**

d. $hist(Boston\$crim[Boston\$crim > 1])$
**From the histogram we can find that 18 suburbs has high crime rate.Here we are considering**

$hist(Boston\$tax, xlim = c(0, 900), ylim = c(0, 150))$
**From the histogram we can find that almost 150 suburbs has tax rate over 650.**



$hist(Boston\$ptratio)$
**From the histogram we can find 200 suburbs has pupil teacher ration over 20.**



e. $dim(subset(Boston, chas == 1))$
35 suburbs bound charles river.

f. $median(Boston\$ptratio)$
**Median pupil teacher ratio is 19.05.**

g. $which.min(Boston\$medv)$
**[1] 399. Therefore the suburb with lowest median value of owner occupied home is #399.**

t(subset(Boston, medv == min(Boston$medv)))

399 406

crim 38.3518 67.9208

**above 3rd quartile**

indus 18.1000 18.1000

**at 3rd quartile**

chas 0.0000 0.0000

**not bounded by river**

nox 0.6930 0.6930

**above 3rd quartile**

rm 5.4530 5.6830

**below 1st quartile**

age 100.0000 100.0000

**maximum**

dis 1.4896 1.4254

**below 1st quartile**
rad 24.0000 24.0000
**maximum**
tax 666.0000 666.0000
**at 3rd quartile**
ptratio 20.2000 20.2000
**at 3rd quartile**
black 396.9000 384.9700
**above 1st quartile**
lstat 30.5900 22.9800
**above 3rd quartile**
medv 5.0000 5.0000
**minimum**

h. $dim(subset(Boston, rm > 7))$
There are 64 suburbs that has more than 7 rooms per dwellings.

$dim(subset(Boston, rm > 8))$
There are 13 suburbs that has more than 7 rooms per dwellings.

If we compare the $Summary(Boston)$ with $dim(subset(Boston, rm > 8))$ then we can find mean crime rate is 0.71879 where as crime rate of Boston is 3.61352.

4. **Definition :**
Multicollinearity is defined as the situation when collinearity is present between three or more variables even if no pair of variables has particularly high correlation between them.

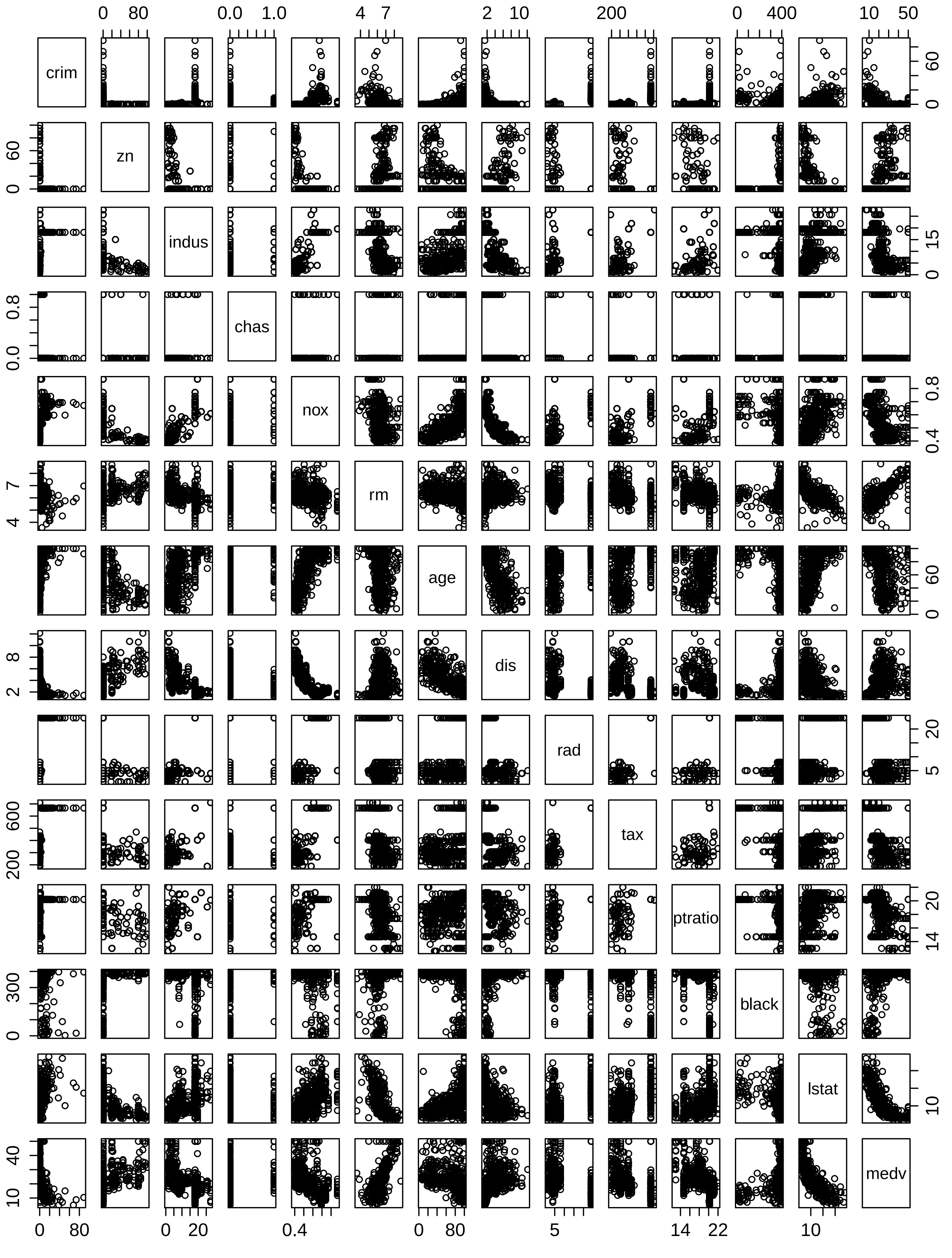
   **Detection :**
We can detect multicollinearity by calculating the variance inflation factor or VIF . VIF measures how much the variance of the estimated regression coefficients are inflated as compared to when the predictor variables are not linearly related.The smallest posible value of VIF is 1, which indicates complete absence of collinearity.If the VIF value goes over 5 or 10 then it indicates large amount of collinearity among the predictor variables.

5. Yes, we can trust the summary output.

   From the given two figures first we need to find out if the assumptions of linear regression are holding or not.The assumptions are,
$E[e_i] = 0, Var[e_i] = \sigma^2, Cov(e_i, e_j) = 0$
Now from the given figures we can find that $E[e - i] = 0$ and $Var[e - i] = \sigma^2$ is holding true.But we can not be certain that $Cov(e_i, e_j = 0$ or not.

   We can perform hypothesis test $H_0$ and $H_a$ to strengthen our conclusion regarding the summary output.

From the given R output we can see that standard error is smaller compared to the coefficient estimates. Consequently we also have high t-value and significantly low p-value. Therefore we can conclude that $H_0$ is not true.
As we have just found that there is association between predictor variable and response variable we need find out the extent of association with the help of $H_a$.Now we find that our $R^2 - Statistics$ is relatively close to 1 and the p-value of F-statistics is significantly low.Therefore we can conclude that the quadratic regression model is a good fit.


crim
zn
indus
chas
nox
rm
age
dis
rad
tax
ptratio
black
lstat
medv
0
80
0.0
1.0
4
7
2
10
200
0
400
10
50
60
0
60
0
15
0
0.8
0.0
0.8
0.4
7
4
60
0
8
2
20
5
600
400
200
20
14
300
0
10
40
10
0
80
0
20
0.4
0
80
5
14
22
10