Assignment 3

Terry Liu 刘骏杰 1630005038

## Question One

The datat file **Real-Estate.txt** contains information on the homes sold in the Denver area during the yea 2003. The variables in this data file are as follows:

|  |  |
| --- | --- |
| Name | Representation |
| Price | Selling price in $1000 |
| Bedrooms | Number of bedrooms |
| Size | Size of the home in square feet |
| Pool | Swimming Pool(1=Yes,0=No) |
| Distance | Distance from the home to the center of the city |
| Township | Township NO. |
| Garage | Garage attached(1=Yes,0=No) |
| Baths | Number of bathrooms |

### Part A

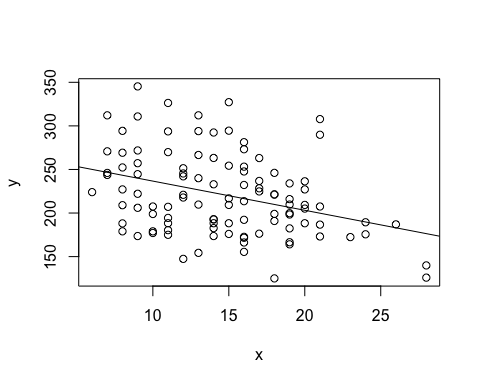
Give a plot of selling price agaignst distance from the home to the center of the city. Does there seem to be a relationship between the two variables? If so, is the relationship direct or inverse ?

### Answer:

data1 <- read.table("Real-Estate.txt",header=TRUE)  
names(data1)

## [1] "Price" "Bedrooms" "Size" "Pool" "Distance" "Township"  
## [7] "Garage" "Baths"

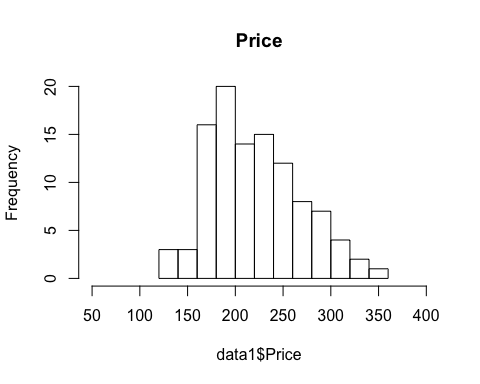
x <- data1$Distance  
y <- data1$Price  
  
plot(x,y)  
abline(lm(y~x))

 We can have the conclusion that there is negative relationship between distance and the selling prze.

### Part B

Plot a histogram for selling price. Comment on the shape of the distribution.

hist.default(data1$Price,xlim=c(50,400),main='Price')



skewness <- function (x, na.rm = FALSE)   
  
{  
  
 if (na.rm)   
  
 x <- x[!is.na(x)]  
  
 sum((x - mean(x))^3)/(length(x) \* sd(x)^3)  
  
}  
  
  
skewness(data1$Price)

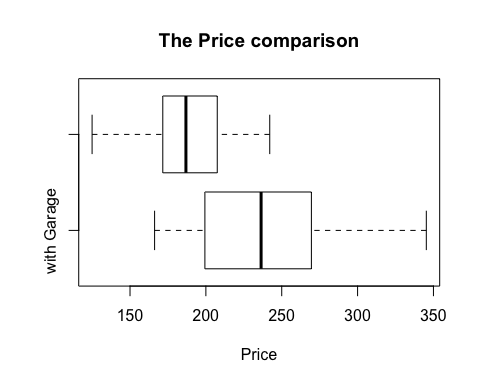
## [1] 0.4605562

We can get the skewness of these data (Selling Price) is equals to , so, it is Positive Skew.

### Part C

Create a side-by-side box plots to compare the distribution of selling price for homes with an attached garage and homes without a garage. Comment on what you observe.

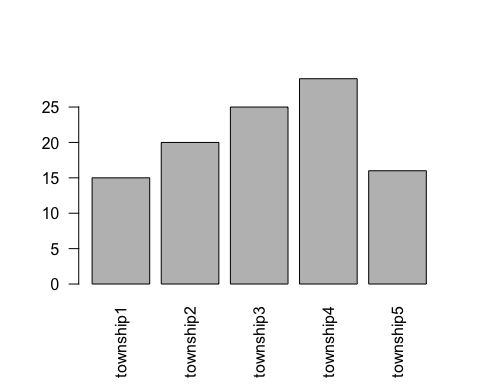
x\_1 <- data1$Price[data1$Garage==1]  
x\_2 <- data1$Price[data1$Garage==0]  
Garage <- c("with Garage","without Garage")  
boxplot(x\_1,x\_2,names = Garage, horizontal = TRUE, main = "The Price comparison",xlab="Price")

 Observation: The averge price of those have garage are higher than without garage.

### Part D

There are five Townships in htis data. Draw a bar plot to compare the number of homes sold in these Townships.

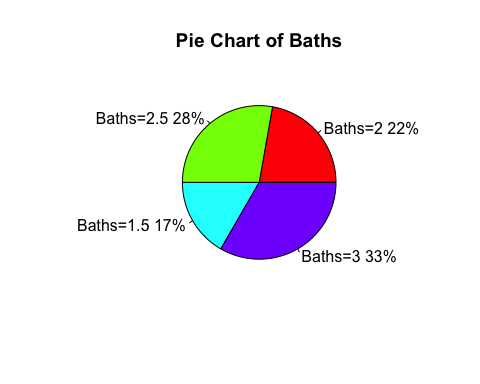
township\_1 <- length(data1$Township[data1$Township == 1])  
township\_2 <- length(data1$Township[data1$Township == 2])  
township\_3 <- length(data1$Township[data1$Township == 3])  
township\_4 <- length(data1$Township[data1$Township == 4])  
township\_5 <- length(data1$Township[data1$Township == 5])  
  
name\_township <- c("township1", "township2", "township3", "township4", "township5")  
township <- c(township\_1, township\_2, township\_3 ,township\_4, township\_5)  
tmp <- as.vector(name\_township)   
  
bar <- barplot(township, axes=T, names = name\_township, las=2)



### Part E

Draw a pie chart of the varialbe “Baths” and label the percentage of each number of Bathrooms

x <- c(unique(data1$Baths))  
  
length\_x <- length(x)  
  
  
  
lbls <- x  
pct <- round(x/sum(x)\*100)  
lbls <- paste(lbls, pct) # add percents to labels   
lbls <- paste("Baths=",lbls,"%",sep="") # ad % to labels   
pie(x,labels = lbls, col=rainbow(length(lbls)),  
 main="Pie Chart of Baths")



## Question Two

Recently, it has been proposed that there is a non-linear relation between the number of perople in a city, and its economic output per person (per captia metropolitan product):

Using data from the 366 metropolitan area of the US (as of 2006) to estimate , by minimizing the mean squared error,

over all a. We will ess later in the class how we could estimate both parameters at onece, but for now we’ll treat as know, and in particular equal = .

### Part A

Using the read.table command, load the gmp.dat (on Ispace) into a data frame called gmp.

data2 <- read.table('gmp.dat',header=T)

#### Part B

The last two columns give, for each city, the total gross metropolitan product () and the per-capita metropolitan product (). Use this to make a vector which has the population () for each city.

y\_0 <- 3611  
y\_i <- as.vector(data2$pcgmp)  
yN\_i <- as.vector(data2$gmp)  
  
N\_i <- yN\_i /y\_i  
  
N\_i <- as.vector(N\_i)

#### Part C

Add the vector of populations to the gmp data fram as a new column, named pop.

data2$pop <- N\_i  
  
names(data2)

## [1] "MSA" "gmp" "pcgmp" "pop"

#### Part D

Calcluate the mean squared error (MSE) of the power-law model, when (in dollars) and

y\_0 <- 3611  
a <- 0.15  
  
MSE <- 1 / 366\* sum((y\_i - y\_0 \* N\_i ^ a) ^ 2)   
MSE <- sprintf("%.2f",MSE)  
  
paste("The MSE equals",MSE)

## [1] "The MSE equals 135420122.48"

#### Part E

Make a sequence of values of a, from to , in steps of . Calculate the MSE for each value of in that sequence, and store all the MSEs in a vector.

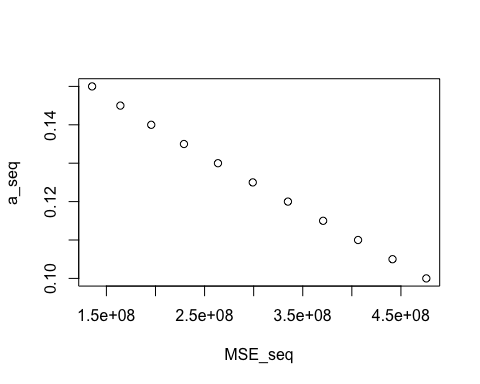
a\_seq <- seq(0.1, 0.15, by=0.005)  
MSE\_seq <- NULL  
  
for (i in 1:length(a\_seq)){  
 MSE\_seq[i] <- 1 / 366\* sum((y\_i - y\_0 \* N\_i ^ a\_seq[i]) ^ 2)  
}  
print(MSE\_seq)

## [1] 475861591 441435384 406335543 370747673 334903031 299086418 263645322  
## [8] 229000490 195658147 164224115 135420122

#### Part F

Plot the square roots of the MSEs vs. a. What are the units of the root mean squared error? What value of a, among those you looked at, leads to the least squared error?

plot(MSE\_seq,a\_seq)



### Question Three

Using R to answer the following questions:

#### Part A

For normal distribution , find a number such that

x <- seq(-10,10,by=0.0000001)  
norm\_x <- pnorm(x,mean = 0, sd = 1)  
norm\_x\_inverse <- pnorm(-x, mean = 0, sd = 1)  
  
y <- abs(norm\_x - norm\_x\_inverse)  
  
for( i in 1:length(y)){  
 if(y[i] == 0.05){  
 print(x[i])  
 }  
}

#### Part B

How much area (probability) is to the right of for a normal distribution ?

1 - pnorm(1.2,mean = 0, sd = 2)

## [1] 0.2742531

### Question Four

Calculate the probability for each of the following events:  
#### Part A

A normally distributed variable with mean and standard deviation is larger than .

1 - pnorm(42, mean = 35, sd = 6)

## [1] 0.1216725

Getting out of successes in a binomial distribution with probability .

choose(10,7) \* 0.8 ^ 7 \* 0.2 ^ 3

## [1] 0.2013266

### Question Five

Three fair dice are rolled, and all outcomes are equally likely. Write a simulation program using R to estimate the probability that the sum of all the numbers on the dice is at least .

dice <- function(x){  
 time1 <- Sys.time()  
 dice\_1 <- sample(1:6, x, replace=T)  
 dice\_2 <- sample(1:6, x, replace=T)  
 dice\_3 <- sample(1:6, x, replace=T)  
 dice\_total <- dice\_1 + dice\_2 + dice\_3   
 SUM\_10 <- dice\_total[dice\_total >= 10]  
 PRO <- length(SUM\_10) / length(dice\_total)  
 time2 <- Sys.time()  
 time\_t <- time2 - time1  
 print(time\_t)  
 return(PRO)  
   
 }  
  
# function test  
dice(10)

## Time difference of 0.0001208782 secs

## [1] 0.8

### Question Six

The file Esti.txt in Ispace contains a set of data which consists of observations.

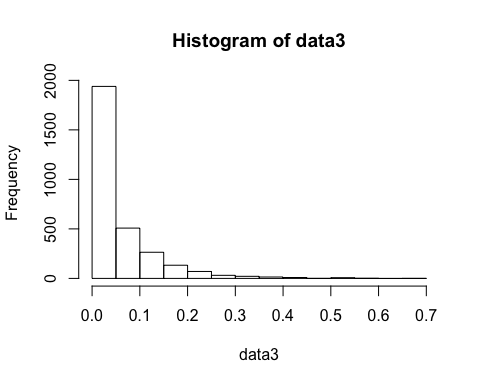
#### Part A

Make a histogram for the data. Which distribution would be a plausible model?

data3 <- read.table("Esti.txt", header = FALSE)

## Warning in read.table("Esti.txt", header = FALSE): incomplete final line  
## found by readTableHeader on 'Esti.txt'

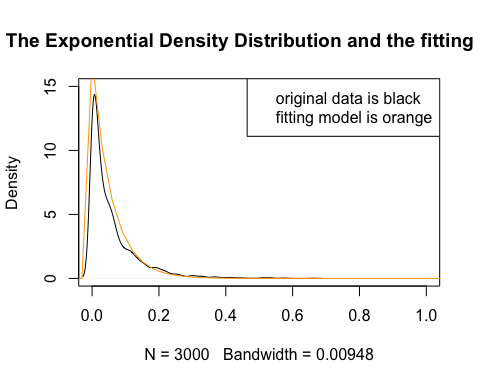
data3 <- as.numeric(data3)  
hist(data3)

 According to the histogram, I think Exponential distribution model would be a plausible model.

#### Part B

Suppose that the exponential distribution would be used to fit the data. Use R to fit the parameters by the method of moments and maximum likelihood.

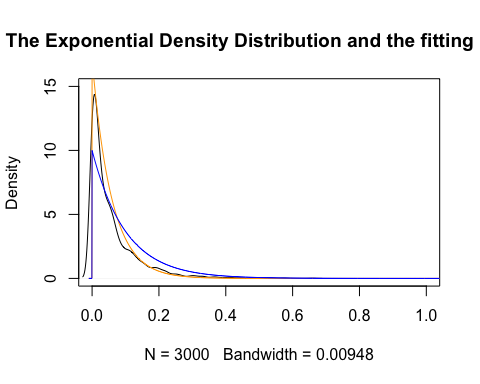
#   
lambda <- 1 / mean(data3)  
den1 <- density(data3)  
plot(den1, xlim=c(0,1), ylim = c(0,15),main="The Exponential Density Distribution and the fitting line")  
set.seed(1)  
x<-seq(-1,2,length.out=100)  
lines(x,dexp(x,lambda),col="orange")  
  
legend("topright",legend=paste(" " = c("original data is","fitting model is"),col=c("black","orange")))



# legend("topright",legend=paste("data", c("original data","fitting model"), lwd=1,col=c("black","orange"))

If we use the MLE:

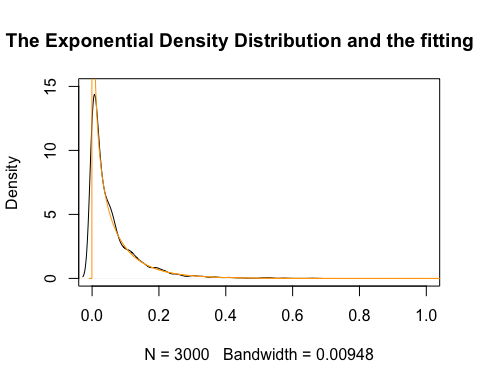
y<- data3  
  
nllik <- function (lambda, obs) - sum(dexp(obs, lambda, log = TRUE))  
  
lambda2 <- optimize(nllik, lower = 0, upper = 10, obs = y)$minimum  
  
plot(den1, xlim=c(0,1), ylim = c(0,15),main="The Exponential Density Distribution and the fitting line")  
set.seed(1)  
x<-seq(-0.01,2,length.out=3000)  
lines(x,dexp(x,lambda),col="orange")  
lines(x,dexp(x,lambda2), col="blue")



#### Part C

Suppose that the gamma distribution is also used to fit the data. Use R to fit the parameters by the method of moments.

gamma\_MME <- function (x)   
{  
 return(c(shape = (mean(as.matrix(x), na.rm = T)/apply(as.matrix(x),   
 2, sd, na.rm = T))^2, scale = apply(as.matrix(x), 2,   
 var, na.rm = T)/mean(as.matrix(x), na.rm = T)))  
}  
  
gammaMME <- gamma\_MME(data3)  
gammaMME <- as.vector(gammaMME)  
  
plot(den1, xlim=c(0,1), ylim = c(0,15),main="The Exponential Density Distribution and the fitting line")  
lines(x,dgamma(x,shape = gammaMME[1],scale = gammaMME[2]),col="orange")



#### Part D

Compare the two estimates obtained by (2) and (3). Which estimation is better?

I have no idea.