**Business Problem Understanding**

A nationwide survey of hospital costs conducted by the US Agency for Healthcare consists of hospital records of inpatient samples. The given data is restricted to the city of Wisconsin and relates to patients in the age group 0-17 years. The agency wants to analyze the data to research on the healthcare costs and their utilization.

Step 1:

Data Collection

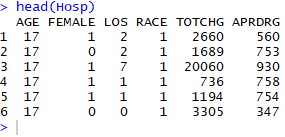
getwd()

Set working directory

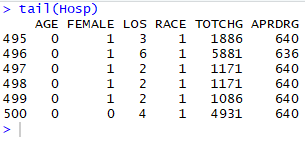
setwd("C:/Users/Ansari/Documents/Projects")

Hosp<-read.csv("HospitalCosts.csv, header = TRUE")

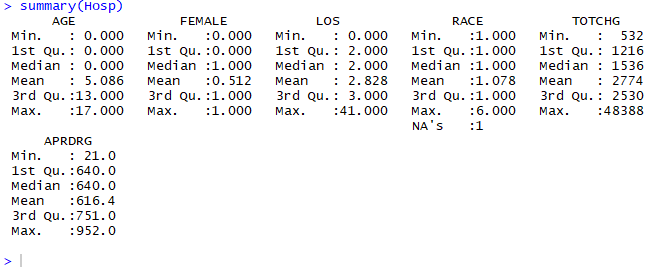
head(Hosp)



tail (Hosp)



summary(Hosp)

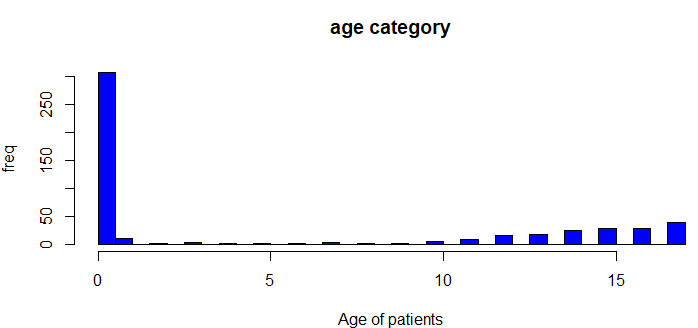


*Question 1:*

To record the patient statistics, the agency wants to find the age category of people who frequent the hospital and has the maximum expenditure.

Attach(Hosp)

hist(AGE)



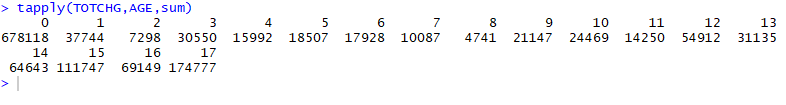
Age<-as.factor(Hosp$AGE)

summary(as.factor(Hosp$AGE)



*Results – The category of 0 (children) appears to be the most frequent visitors to the hospital.*

tapply(TOTCHG,AGE,sum) ##sum of hospital cost for each age



max(tapply(TOTCHG,AGE,sum)) ##maximum cost



which.max(tapply(TOTCHG,AGE,sum))



*Results – max expenditure by age 0 = 678118, 15=111747 and 17 = 174777.*

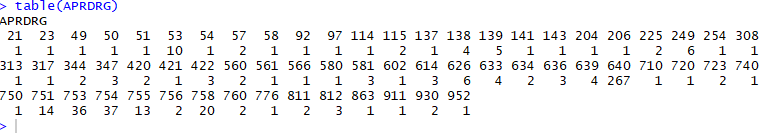
\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

*Question 2:*

In order of severity of the diagnosis and treatments and to find out the expensive treatments, the agency wants to find the diagnosis related group that has maximum hospitalization and expenditure

APRDRG

table(APRDRG)



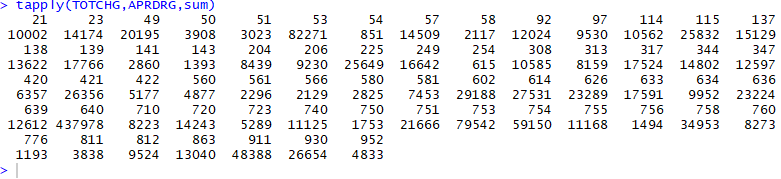
max(table(APRDRG))



which.max(table(APRDRG))



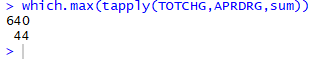
tapply(TOTCHG,APRDRG,sum)



max(tapply(TOTCHG,APRDRG,sum))



which.max(tapply(TOTCHG,APRDRG,sum))



*Results – The group 640 has the maximum number of expenditure and the highest total of hospilazation costs of - 437978.*

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

*Question 3:*

To make sure that there is no malpractice, the agency needs to analyze if the race of the patient is related to the hospitalization costs.

Null hypothesis -

Ho = Race has no effect on Hospitalization Cost

H1 = Race has effect on hospitalization Cost

Dependent Variable (DV) = Hosp Cost -> Num

Independent Variable (IV) = Race -> Char

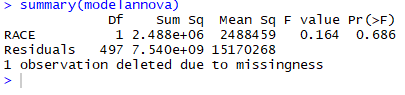
Hypothesis test -> Anova test

p<0.05 -> IV has significant effect on your DV

Race<-as.factor(RACE)

anova\_output <- aov(TOTCHG~RACE, data=dataset)

summary(anova\_output)



*Results – pvalue is very high at 68% so we can take a risk and reject the null hypothesis. This means that there is no relation between race and hospital costs.*

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

*Question* *4:*

To properly utilize the costs, the agency has to analyze the severity of the hospital costs by age and gender for proper allocation of resources.

Dependent Variable = Hosp Cost

Independent Variable = Age , Gender

# pvalue(Age) < 0.05 -> Age is important factor in the hospital costs

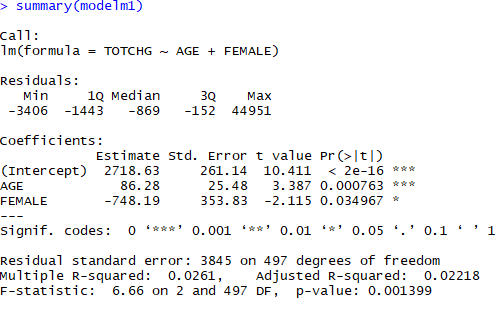
# pvalue(Gender) < 0.05 -> Gender is also having the impact on cost

# To properly utilize the costs, the agency has to analyze the severity of

# the hospital costs by age and gender for proper allocation of resources.

modelm1<-lm(TOTCHG~AGE+FEMALE)

summary(modelm1)



*Results – p-value for age is very less which means that it’s a important factor in hospital costs as seen by the levels in p-values.*

*Gender is also less p-value which also means that there is an impact on cost .*

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

*Question 5:*

Since the length of stay is the crucial factor for inpatients, the agency wants to find if the length of stay can be predicted from age, gender, and race.

Dependent Variable = los

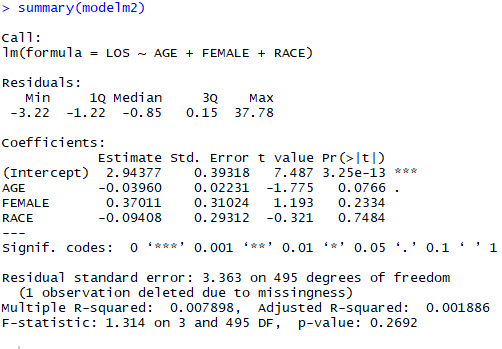
Independent Variable = age, gender, and race

p<0.05 ->Reject Null hypothesis

p>0.05 ->Accept Null hypothesis

modelm2<-lm(LOS~AGE+FEMALE+RACE)

summary(modelm2)



*Results – p-value indicates that there is no linear relation between the three variables – Age, Gender and Race.*

*So we cannot predict the stay (Los) of the patients.*

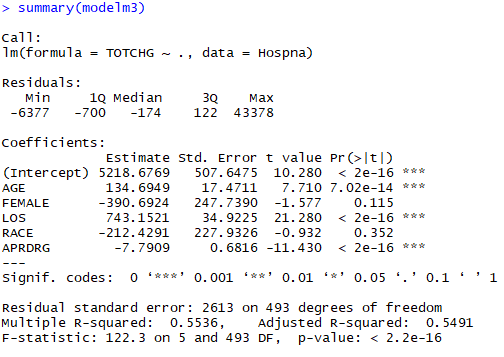
\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

*Question 6:*

To perform a complete analysis, the agency wants to find the variable that mainly affects the hospital costs.

modelm3<-lm (TOTCHG~ .,data=hospna)

summary(modelm3)



*Results – We can see that the length of the stay (Los) affects the total hospital costs.*