**BUSINESS SCENARIO**

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.

Attributes Description

AGE : Age of the patient discharged

FEMALE : Binary variable that indicates if the patient is female

LOS : Length of stay, in days

RACE : Race of the patient (specified numerically)

TOTCHG : Hospital discharge costs

APRDRG : All Patient Refined Diagnosis Related Groups

The data has been downloaded from the URL mentioned below (under the name HospitalCosts):

[http://instruction.bus.wisc.edu/jfrees/jfreesbooks/Regression%20Modeling](http://instruction.bus.wisc.edu/jfrees/jfreesbooks/Regression%20Modeling/BookWebDec2010/data.html)

[/BookWebDec2010/data.html](http://instruction.bus.wisc.edu/jfrees/jfreesbooks/Regression%20Modeling/BookWebDec2010/data.html)

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.**

**CODE:**

hospital<-read.csv(file.choose(),header = TRUE)

str(hospital)

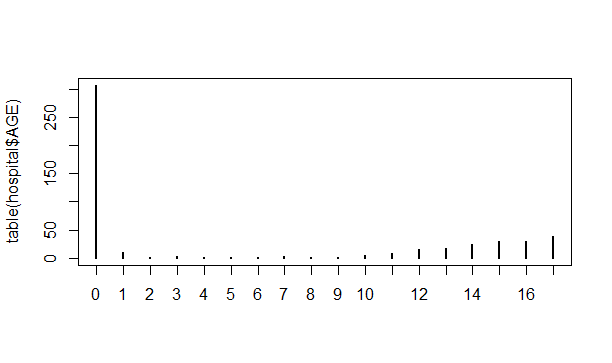
table(hospital$AGE)

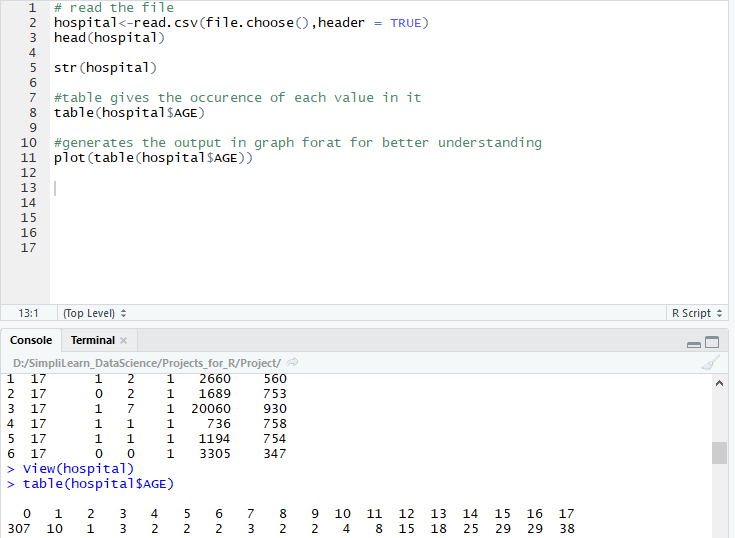
plot(table(hospital$AGE))

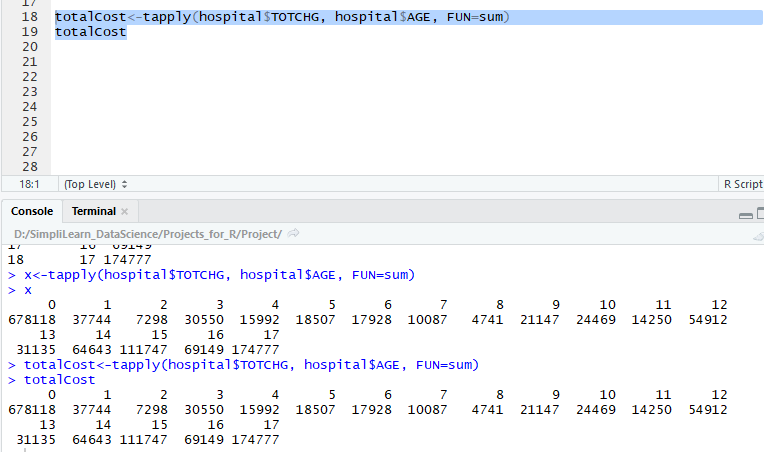
totalCost<-tapply(hospital$TOTCHG, hospital$AGE, FUN=sum)

totalCost

**OUTPUT:**







**RESULT:**

As we can see from the output that the

Maximum Hospital Visit=**307**

Maximum Expenditure =**678118**

1. **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.**

**CODE:**

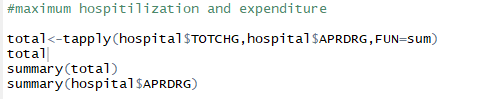
total<-tapply(hospital$TOTCHG,hospital$APRDRG,FUN=sum)

total

summary(total)

summary(hospital$APRDRG)

**OUTPUT:**







**RESULT:**

As we can see fro the output that the category **640** has the maximum number of entries of hospitilization and also has the highest hospitilization cost i.e **437978**

**---------------------------------------------------------------------------------------------------------------------------------------**

1. **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.**

**CODE:**

str(hospital)

hospitalRace<-as.factor(hospital$RACE)

summary(hospitalRace)

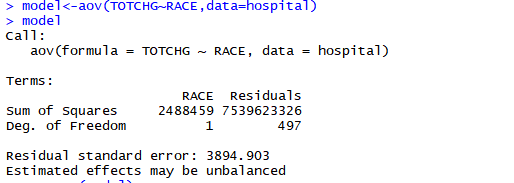
any(is.na(hospitalRace))

hospitalRace<-na.omit(hospitalRace)

model<-aov(TOTCHG~RACE,data=hospital)

model

**OUTPUT:**



summary(model)



table(hospital$RACE)



**RESULT:**

From the output we can observe that the p-value is very high specifying that there is no relation between the race of patient and the hospitization cost.

From the summary also we can see that 484 patients are of Race 1 out of 500 patients.

Hence we can conclude that there is no race wise cost biasness.

**---------------------------------------------------------------------------------------------------------------------------------------**

1. **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.**

From the dataset we can conclude that variables

Dependent variable : TOTCHG

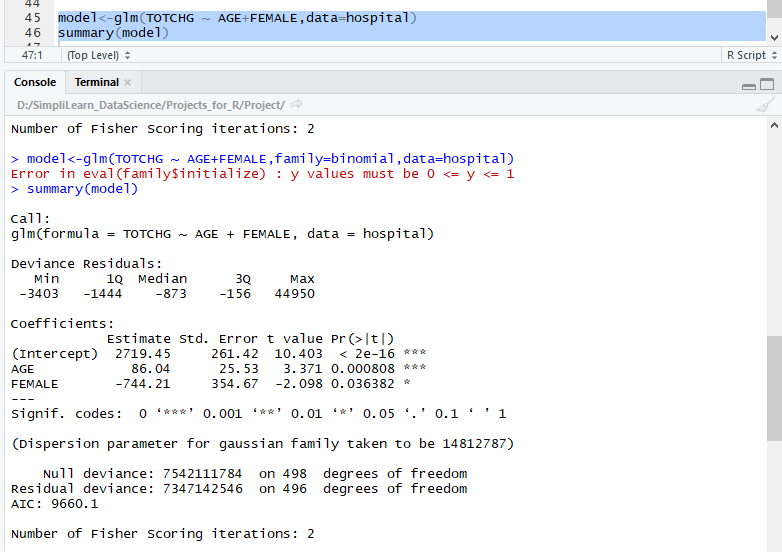
Independent variable : AGE,FEMALE

**CODE:**

model<-glm(TOTCHG ~ AGE+FEMALE,data=hospital)

summary(model)

**OUTPUT:**



**RESULT:**

From the output we can see that AGE is an important factor in hospital cost by looking into the p values.

The gender also seems to have an impact.

Based on the negative coefficients we can conclude that females incur less cost than males.

1. **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.**

From the question we can predict

DEPENDENT VARIABLE - LOS

INDEPENDENT VARIABLE - AGE,FEMALE and RACE

**CODE:**

hospital$RACE<-as.factor(hospital$RACE)

hospital$FEMALE<-as.factor(hospital$FEMALE)

head(hospital)

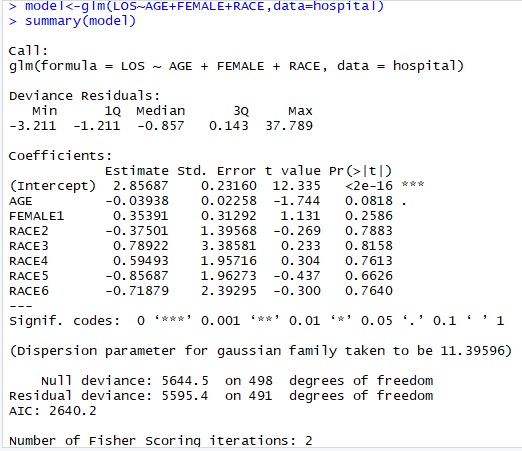
str(hospital)

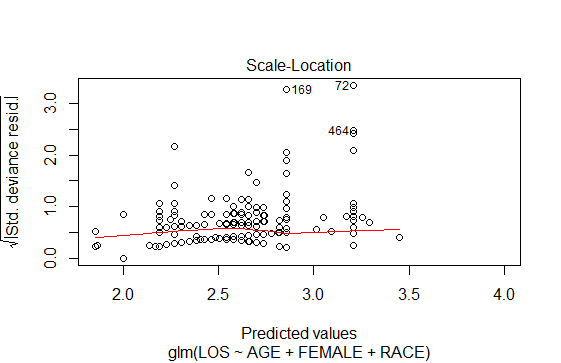
model<-glm(LOS~AGE+FEMALE+RACE,data=hospital)

summary(model)

plot(model)

**OUTPUT:**





**RESULT:**

The significance codes are almost null for all the variables ,except for the intercept

The high p values signifies their is no linear relationship between the given variable

Hence we cannot predict the length of stay of the patients based on age,gender and race

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

We have to predict a model that will affect the hospital cost

So in this case we have one dependent variable and all other independent variable

Dependent Variable - TOTCHG

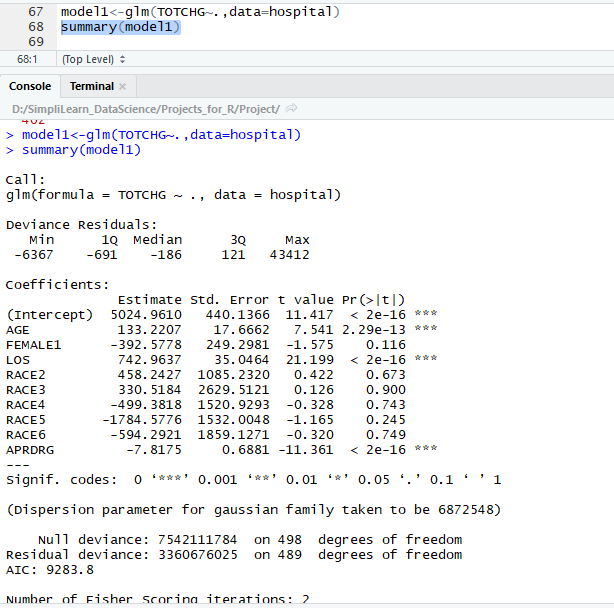
INDEPENDENT VARIABLE - All other variables

**CODE:**

model1<-glm(TOTCHG~.,data=hospital)

summary(model1)

**OUTPUT:**



**RESULT:**

From the output we can see that Age and Length of Stay affects the cost of hospitalization

Cost is directly proportional to the length of stay

As per the output we can see that an increase of 1 day stay.the hospital cost will increase by 742.