Healthcare cost analysis

By Ifalore Simeon

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 analyse the data to research on healthcare costs and their utilization.

Domain: Healthcare

Dataset Description:

Here is a detailed description of the given dataset:

Attribute	Description
Age	Age of the patient discharged
Female	A 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

Analysis to be done:

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

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.

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.

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

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.

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

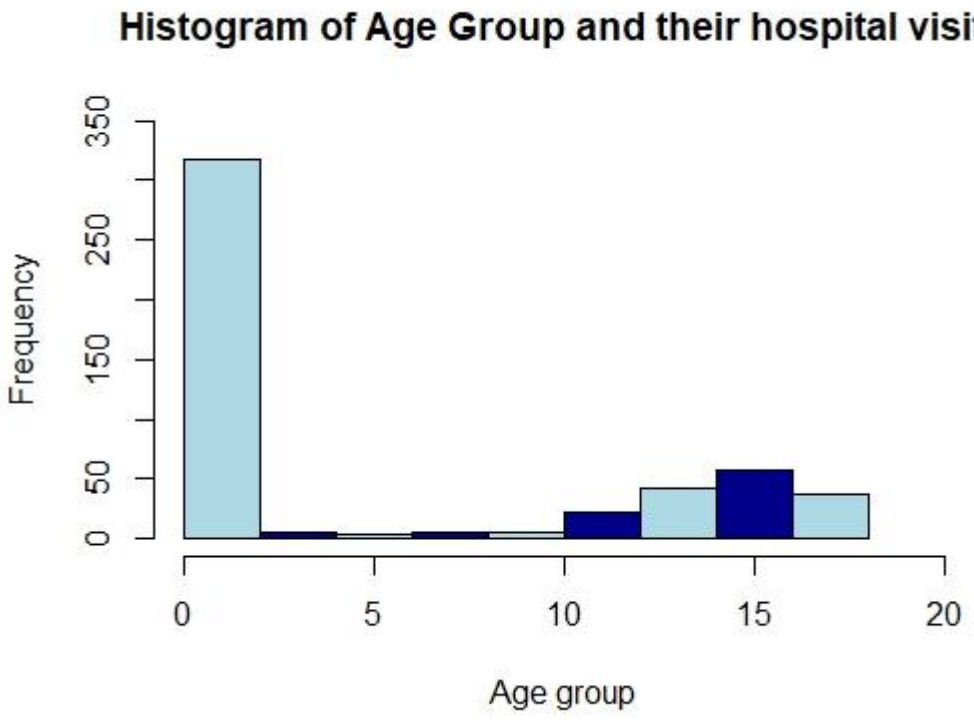
1. To record the patient statistics, the agency wants to find the age category of people who frequently visit the hospital and has the maximum expenditure.

importing the Data set

```
getwd()

setwd("C:/Users/Simeon/desktop")

library(readxl)

HospitalCost <- read_excel("HospitalCost.xlsx")

View(HospitalCost)
```

1. To find the age category of people who frequently visit the hospital and has the maximum expenditure.

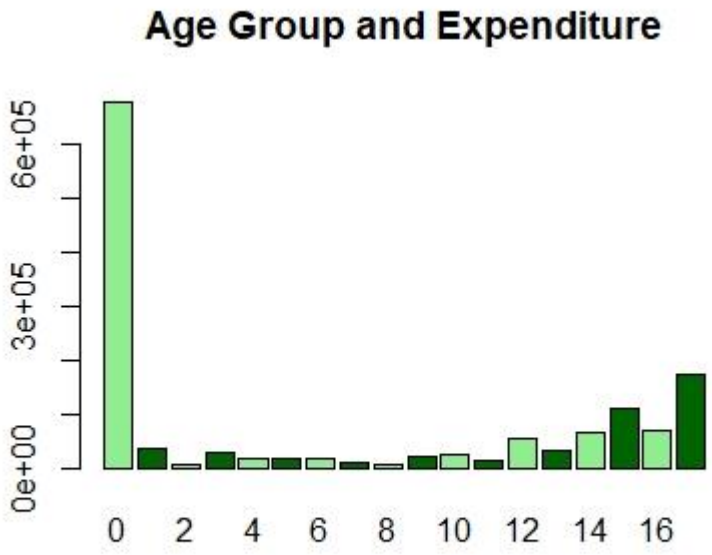
```
summary(as.factor(HospitalCost$AGE))

hist(HospitalCost$AGE, main = "Histogram of Age Group and their hospital visits", xlab = "Age group", border = "black", col = c("light blue", "dark blue"), xlim = c(0,20), ylim = c(0,350))
```



As can be seen here, the maximum number of hospital visits are for age group is 0-1 years

Summarize expenditure based on age group

```
ExpenseBasedOnAge <- aggregate(TOTCHG ~ AGE, FUN=sum, data=HospitalCost)

which.max(tapply(ExpenseBasedOnAge$TOTCHG, ExpenseBasedOnAge$TOTCHG, FUN=sum))

barplot(tapply(ExpenseBasedOnAge$TOTCHG, ExpenseBasedOnAge$AGE, FUN=sum),border =
"black", col = c("light green", "dark green")
```



Maximum expenditure is 678118 for 0-1 year olds.

2. Diagnosis-related group that has maximum hospitalization and expenditure.

```
summary(as.factor(HospitalCost$APRDRG))

DiagnosisCost <- aggregate(TOTCHG ~ APRDRG, FUN = sum, data = HospitalCost)

DiagnosisCost[which.max(DiagnosisCost$TOTCHG), ]
```

Using the code above to get the output below we can see that **640** diagnosis related group had a max cost of **437978**

```
    APRDRG TOTCHG
44     640 437978
```

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.

After checking the RACE column it is noticed that there is an NA variable as part of the data which is removed using the following code

summary(as.factor(HospitalCost$RACE))

HospitalCost <- na.omit(HospitalCost)

Once the NA variable is removed themodel is built using linear regression to see if race has any influence on hospital cost as shown below.

summary(as.factor(HospitalCost$RACE))

raceInflModel <- lm(TOTCHG ~ RACE, data = HospitalCost)

summary(raceInflModel)

```
Call:
lm(formula = TOTCHG ~ RACE, data = HospitalCost)

Residuals:
   Min     1Q Median     3Q    Max
 -2256  -1560  -1227   -258  45600

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2925.7      405.0   7.224 1.92e-12 ***
RACE          -137.3      339.1  -0.405    0.686
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3895 on 497 degrees of freedom
Multiple R-squared:  0.0003299, Adjusted R-squared:  -0.001681
F-statistic: 0.164 on 1 and 497 DF,  p-value: 0.6856
```

From the output we see that P-value is **0.69** which is much higher than **0.5**, Hence, we can say that race doesn’t affect the hospitalization costs or RACE is not statistically significant.

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

Age_Gender <- as.factor((HospitalCost$FEMALE))

summary(Age_Gender)

```
> summary(Age_Gender)
  0   1
244 255
```

We can see that from the summary the gender between male and female is evenly distributed.

Age_GenderInflModel <- lm(formula = TOTCHG ~ AGE + FEMALE, data = HospitalCost)

summary(Age_GenderInflModel)

```
Call:
lm(formula = TOTCHG ~ AGE + FEMALE, data = HospitalCost)

Residuals:
   Min     1Q Median     3Q    Max
 -3403  -1444   -873   -156  44950

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2719.45     261.42  10.403  < 2e-16 ***
AGE            86.04      25.53   3.371 0.000808 ***
FEMALE       -744.21     354.67  -2.098 0.036382 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3849 on 496 degrees of freedom
Multiple R-squared:  0.02585,   Adjusted R-squared:  0.02192
F-statistic: 6.581 on 2 and 496 DF,  p-value: 0.001511
```

From the output for the linear model for age and gender we can see that AGE is 0.00081 which is much lesser than 0.05 and it also has three stars (***) next to it, it means AGE has the most significance statistically. Also, gender is also less than 0.05. gender is (0.036)

Hence, we can conclude that the model is statistically significant

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.

using linear model to find out the influences of RACE, AGE and GENDER on the LENGTH OF STAY with the code below.

Age_Gender_Race_InflModel <- lm(formula = LOS ~ AGE + FEMALE + RACE, data = HospitalCost)

summary(Age_Gender_Race_InflModel)

```
Call:
lm(formula = LOS ~ AGE + FEMALE + RACE, data = HospitalCost)

Residuals:
   Min     1Q Median     3Q    Max
 -3.22  -1.22  -0.85   0.15  37.78

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.94377    0.39318   7.487 3.25e-13 ***
AGE         -0.03960    0.02231  -1.775   0.0766 .
FEMALE       0.37011    0.31024   1.193   0.2334
RACE        -0.09408    0.29312  -0.321   0.7484
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.363 on 495 degrees of freedom
Multiple R-squared:  0.007898,  Adjusted R-squared:  0.001886
F-statistic: 1.314 on 3 and 495 DF,  p-value: 0.2692
```

From the output p-value is higher than 0.05 for age, gender and race, indicating there is no linear relationship between these variables and length of stay. Hence, age, gender and race cannot be used to predict the length of stay of inpatients.

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

Building a final model to determine the variable affects hospital cost. I have included all variables in the linear model to see which has a statistical significance and which doesn’t

HospCostModel <- lm(formula = TOTCHG ~ ., data = HospitalCost)

summary(HospCostModel)

```
Call:
lm(formula = TOTCHG ~ ., data = HospitalCost)

Residuals:
   Min     1Q Median     3Q    Max
 -6377   -700   -174    122  43378

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 5218.6769   507.6475  10.280  < 2e-16 ***
AGE          134.6949    17.4711   7.710 7.02e-14 ***
FEMALE      -390.6924   247.7390  -1.577    0.115
LOS          743.1521    34.9225  21.280  < 2e-16 ***
RACE        -212.4291   227.9326  -0.932    0.352
APRDRG        -7.7909     0.6816 -11.430  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2613 on 493 degrees of freedom
Multiple R-squared:  0.5536,    Adjusted R-squared:  0.5491
F-statistic: 122.3 on 5 and 493 DF,  p-value: < 2.2e-16
```

Using the output from the HospCostModel we can see that Age, Length of stay (LOS) and patient refined diagnosis related groups (APRDRG) have three stars (***) next to it. Therefore, they are the ones with statistical significance. Also, RACE is the least significant.