Healthcare Cost Analysis

Background and Objective:

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 healthcare costs and their utilization.

Import dataset

```
setwd("C:/Learning/Data Analytics/Datasicence with
R/Project/Healthcare - cost analysis")
HospitalCosts=read.csv("hospital_costs.csv", header=TRUE)
head(HospitalCosts)
names(HospitalCosts)

names(HospitalCosts)[1] = "AGE"

names(HospitalCosts)
```

1. Record patient statistics:

The agency wants to find the age category of people who frequently visit the hospital and has the maximum expenditure.

- *Age*: Age of the patient discharged
- *Totchg*: Hospital discharge costs

```
summary(HospitalCosts)
```

Get number of hospital visits based on age

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

- Total number of hospital for 0-1 age group is 307

```
hist(HospitalCosts$AGE, main="Histogram of Age Group and their
hospital visits",
     xlab="Age group", border="black", col=c("light green", "dark
green"), 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=HospitalCosts)
```

Get the maximum expense and its age group

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

```
barplot(tapply(ExpenseBasedOnAge$TOTCHG, ExpenseBasedOnAge$AGE,
FUN=sum))
```

- Maximum expenditure for 0-1 yr is 678118

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

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: All Patient Refined Diagnosis Related Groups
- Totchg: Hospital discharge costs

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

Get the diagnosis-related group and its hospitalization expenditure

```
DiagnosisCost = aggregate(TOTCHG ~ APRDRG, FUN = sum, data =
HospitalCosts)
```

Get the max diagnostic cost

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

- As can be seen here **640** diagnosis related group had a max cost of **437978**

3. Race vs Hospitalization costs

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.

#Ho (Null hypothesis):Independent variable (RACE) is not influencing dependent variable (COSTS) #H0:there is no no correlation among residuals, # p-value = 0.7394 <== this is > 0.5 #i.e. they are independent #in case of regression, we need high p value so that we cannot reject the null

```
summary(as.factor(HospitalCosts$RACE))
```

- There is one null value. This needs to be removed

```
HospitalCosts = na.omit(HospitalCosts)
summary(as.factor(HospitalCosts$RACE))
```

- As can be seen 484 patients out of 499 fall under group 1, showing that the number of observations for 1 category is way higher than others - hence data is skewed. This will only affect the results from linear regression or ANOVA analysis

```
raceInfluence=lm(TOTCHG~ RACE, data=HospitalCosts)
summary(raceInfluence)
```

- pValue is **0.69** it is much higher than *0.5*
- We can say that race doesn’t affect the hospitalization costs

Anaysis using ANOVA

We can also use anova statistical test for estimating how dependent variable, in this case RACE, affects the independent variable, the hospitalization cost

```
raceInfluenceAOV <- aov(TOTCHG ~ RACE, data=HospitalCosts)
raceInfluenceAOV
```

```
summary(raceInfluenceAOV)
```

- The residual variance (deviation from original) (of all other variables) is very high. This implies that there is very little influence from RACE on hospitalization costs
- As can be seen, the degree of freedom (Df) for RACE is 1 and that of residuals is 497 observations
- The F-Value, the test statistic is 0.16 which is much less than 0.5 showing that RACE doesn't affect teh hospitalization cost.
- The Pr(>F), the p_value of 0.69 is high confirming that RACE does not affect hospitalization cost.

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.

```
HospitalCosts$FEMALE <- as.factor(HospitalCosts$FEMALE)
summary(HospitalCosts$FEMALE)
```

- As can be seen here, there is equal distribution of male and female in the group

```
ageGenderInflModel=lm(TOTCHG~ AGE+FEMALE, data=HospitalCosts)
summary(ageGenderInflModel)
```

- Since the pValues of AGE is much lesser than 0.05, the ideal statistical significance level, and it also has three stars (***) next to it, it means AGE has the most statitical significance
- Similarly, gender is also less than 0.05.
- 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.

```
HospitalCostsNA$RACE <- as.factor(HospitalCostsNA$RACE)

ageGenderRaceInflModel=lm(LOS ~ AGE+FEMALE+RACE, data=HospitalCostsNA)
summary(ageGenderRaceInflModel)
```

- The 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. Complete analysis

The agency wants to find the variable that mainly affects hospital costs.

Significance method - build a model using all independent variables vs dependent variable

```
hospitalCostModel=lm(TOTCHG ~ . , data=HospitalCosts)
summary(hospitalCostModel)
```

- As it is apparent from the coefficient values, Age, Length of stay (LOS) and patient refined diagnosis related groups(APRDRG) have three stars (***) next to it. So they are the ones with statistical significance
- Also, RACE is the least significant. build a model after removing RACE

```
hcm1=lm(TOTCHG ~ AGE+FEMALE+LOS+APRDRG , data=HospitalCostsNA)
summary(hcm1)
```

```
hcm2=lm(TOTCHG ~ AGE+LOS+APRDRG , data=HospitalCostsNA)
summary(hcm2)
```

- Since APRDRG has -ve t-value, dropping it.

```
hcm3=lm(TOTCHG ~ AGE+LOS , data=HospitalCostsNA)
summary(hcm3)
```

Comparing Models

Data	Approach	Model Name	Detail	R2	adj R2	std err	R2 - adj R2	p-value
Hospital Costs	Ap1 :significance	hospitalCostModel	signifi, all independent variables	0.554	0.549	2610	0.005	<2e-16
Hospital Costs	Ap1 :significance	hcm1	-RACE	0.553	0.549	2610	0.004	<2e-16
Hospital Costs	Ap1 :significance	hcm2	-RACE -FEMALE (gender)	0.551	0.548	2620	0.003	<2e-16
Hospital Costs	Ap1 :significance	hcm3	AGE + LOS	0.419	0.416	2970	0.003	<2e-16

- Removing Race and gender doesn't change the R2 value. It doesn't impact cost
- Removing APRDRG in model hcm3 increases the standard error. Hence model hcm2 seems to be better.

Analysis Conclusion:

- As is evident in the multiple models above, health care costs is dependent on age, length of stay and the diagnosis type.

1. Healthcare cost is the most for patients in the 0-1 yrs age group category

i) Maximum expenditure for 0-1 yr is *678118*

2. Length of Stay increases the hospital cost

3. All Patient Refined Diagnosis Related Groups also affects healthcare costs

i) 640 diagnosis related group had a max cost of 437978

4. Race or gender doesn't have that much impact on hospital cost