STA 6244 Data Analysis I

Fall-2022

Healthcare Cost Analysis

Report

For:

Dr. Zhongxue Chen

Presented By:

Kadali Divya

Yaswanth Pothineni

Abstract:

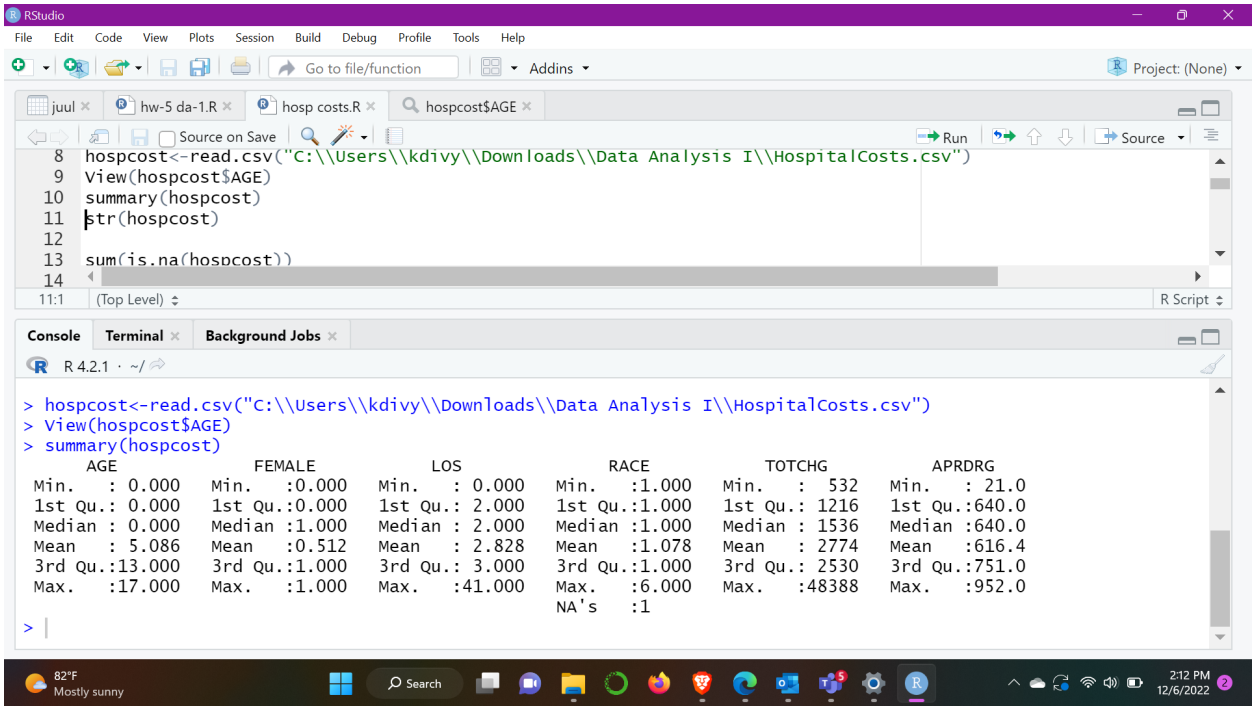
Hospital records of inpatient samples are used in a countrywide study of hospital expenses. The provided information only applies to patients between the ages of 0 and 17 and is specific to the city of Wisconsin. And we want to research healthcare consumption and expenditures,and to evaluate the data. We have also determined the age group of patients who spend the most money and frequently visit the hospital.The dataset we chose has details of hospital costs from getting admitted, taking tests and getting discharged of various age groups.

Introduction:

Health is wealth, being fit is a boon and when unwell, getting well can get exorbitant.Health is wealth, being fit is a boon and when unwell, getting well can get exorbitant. We wanted to record the patient statistics & find the age category of people who frequent the hospital and have the maximum expenditure. Analyzing the severity of the hospital costs by age and gender for proper allocation of resources can get trivial so we wanted to check these attributes while also finding the variable that mainly affects the hospital costs.

Data Sources :

1. Kaggle- Hospital Cost Analysis.
    - Age: Age of the patient who was released
    - FEMALE: A binary value indicating whether the patient is a female
    - LOS: Length of stay, expressed in days
    - RACE: The patient's race (specified numerically)
    - TOTCHG: Charges for leaving a hospital
    - APRDRG : All Patient Refined Diagnosis Related Groups



2.Tools used :

- R Studio

3. Model used :

- Linear Modelling

4.Tests Applied :

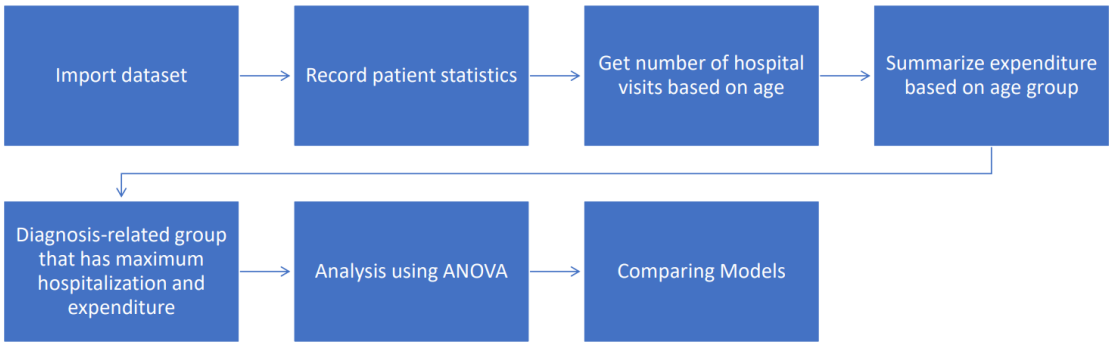
- ANOVA, t test.

Research Questions:

1. Which age group visits the hospital most frequently from the dataset?
2. What factors are dependent on health-care costs?
3. Are there any independent factors that contribute to hospital bills?
4. Is race or age any factor for expenditure?
5. Which kind of test gives out the right analysis for our dataset?

Data Collection:

Methodology



Step #1 Importing data from csv file

```
hospcost<-read.csv("C:\\Users\\kdivy\\Downloads\\Data Analysis I\\HospitalCosts.csv")
View(hospcost$AGE)
summary(hospcost)
str(hospcost)
```

Step #2 Cleaning of Data and analyzing age-wise figures.

```
sum(is.na(hospcost))
dim(hospcost)
hospcost <- na.omit(hospcost)
summary(hospcost)
dim(hospcost)
```

```
hist(hospcost$AGE,xlab = "Age", ylab = 'No. of Visits',col=c("light green", "dark green"),main =
"Age wise Frequency of Patients")
table(hospcost$AGE)
max(summary(as.factor(hospcost$AGE)))
```

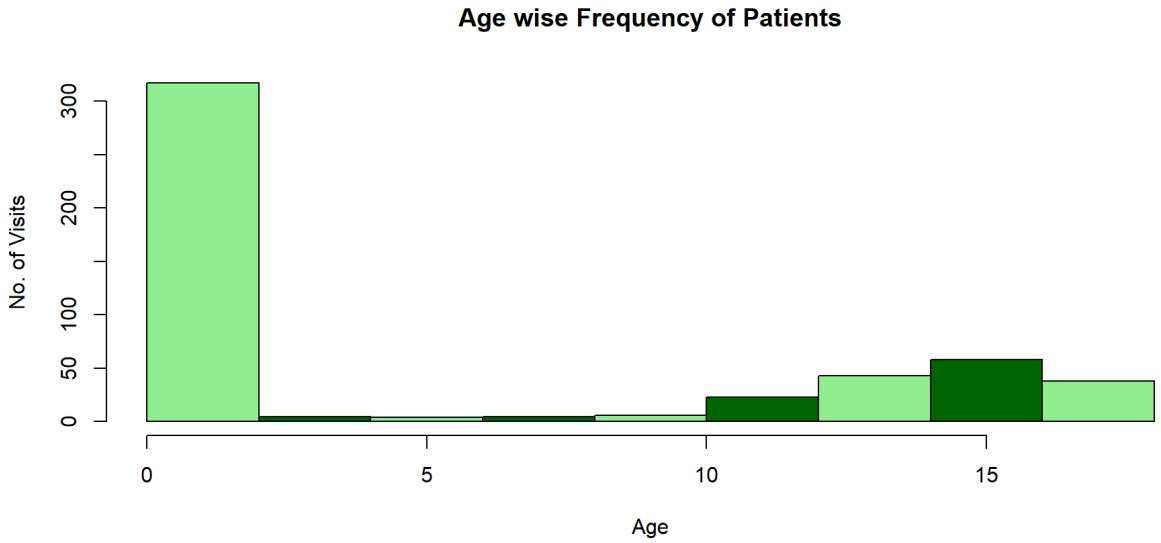


Figure 1 : Age wise number of patients

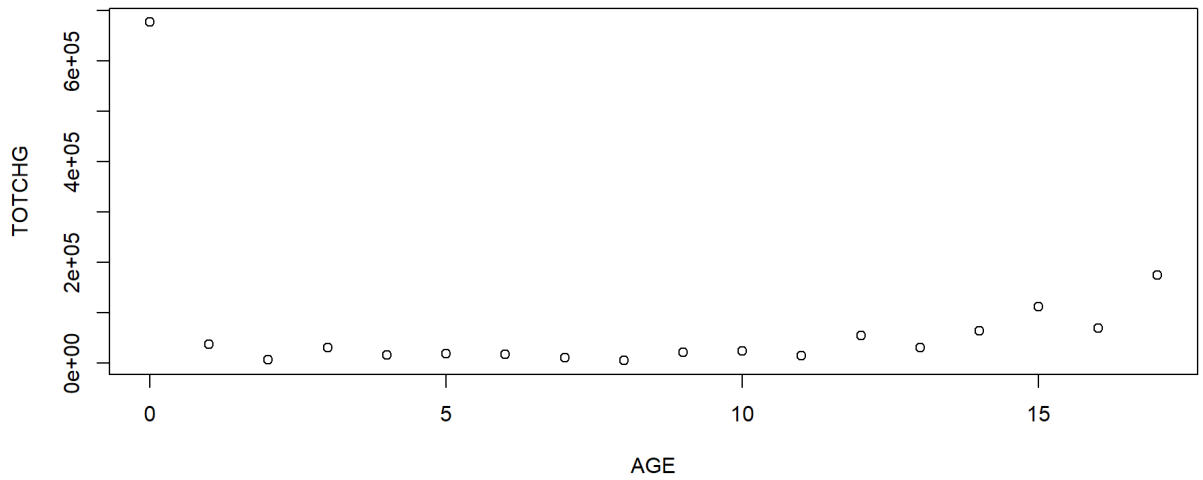
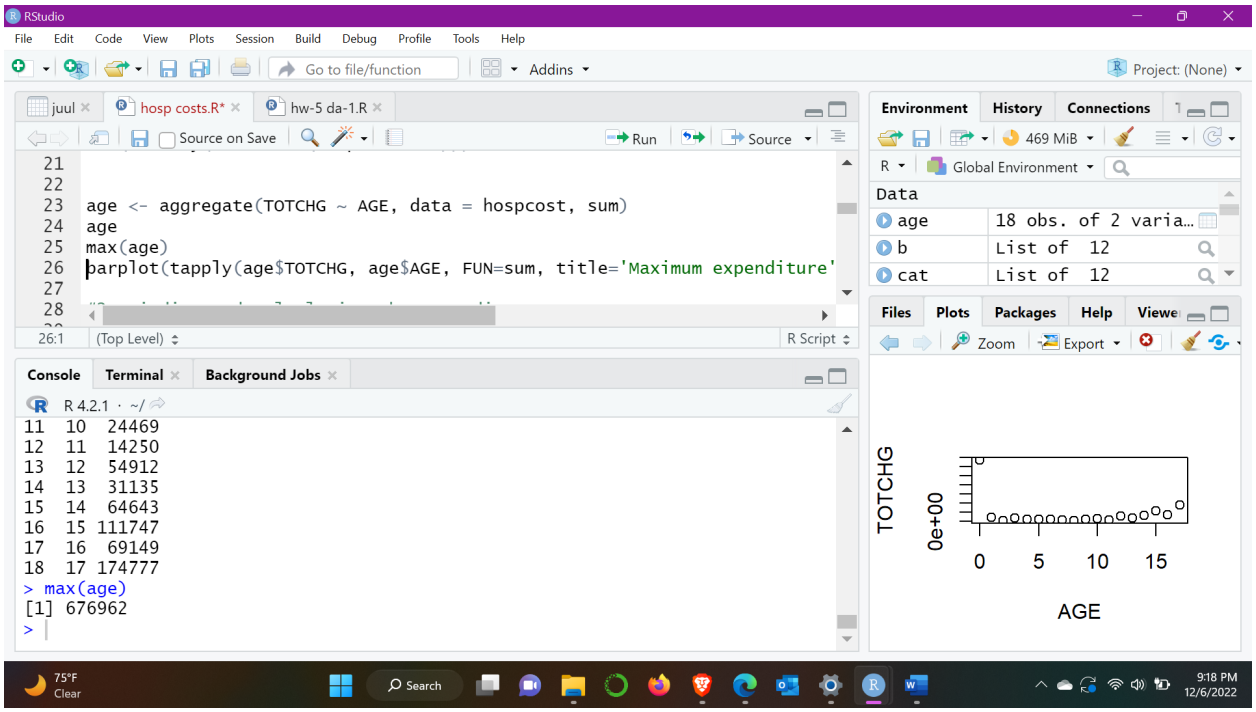
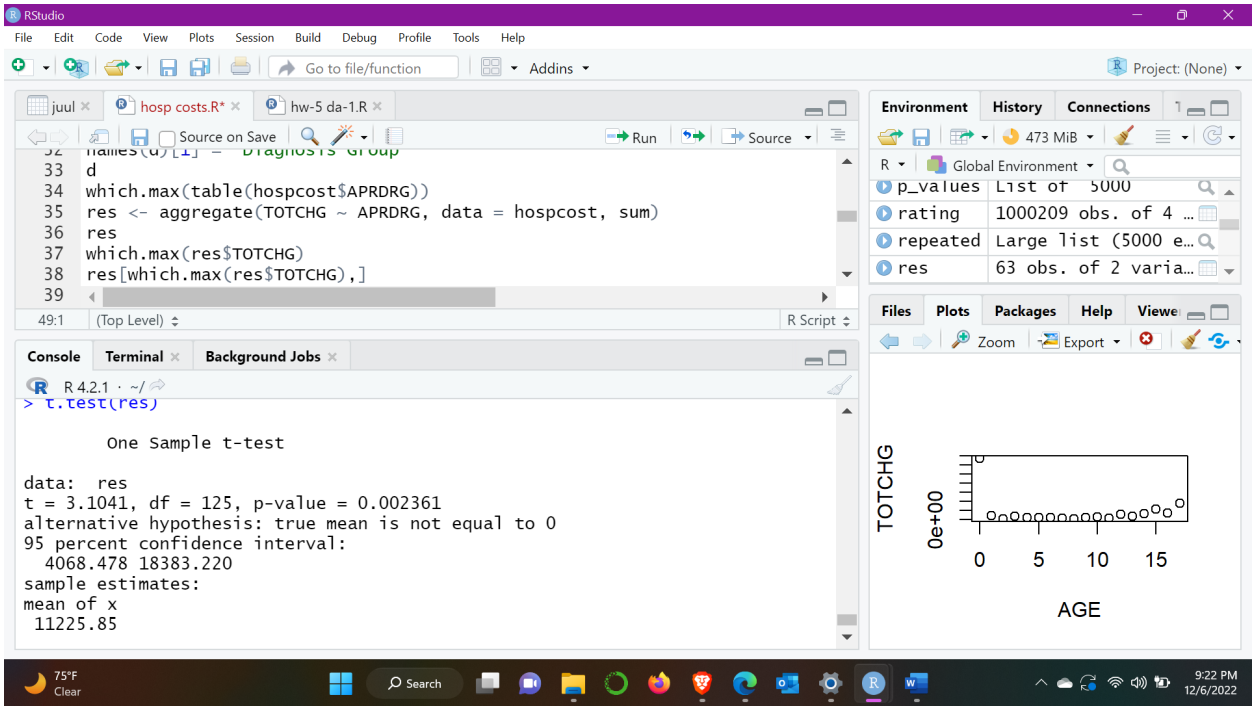


Figure 2 : Number of days every age-group stays

Step #3 Calculating the expenditure :



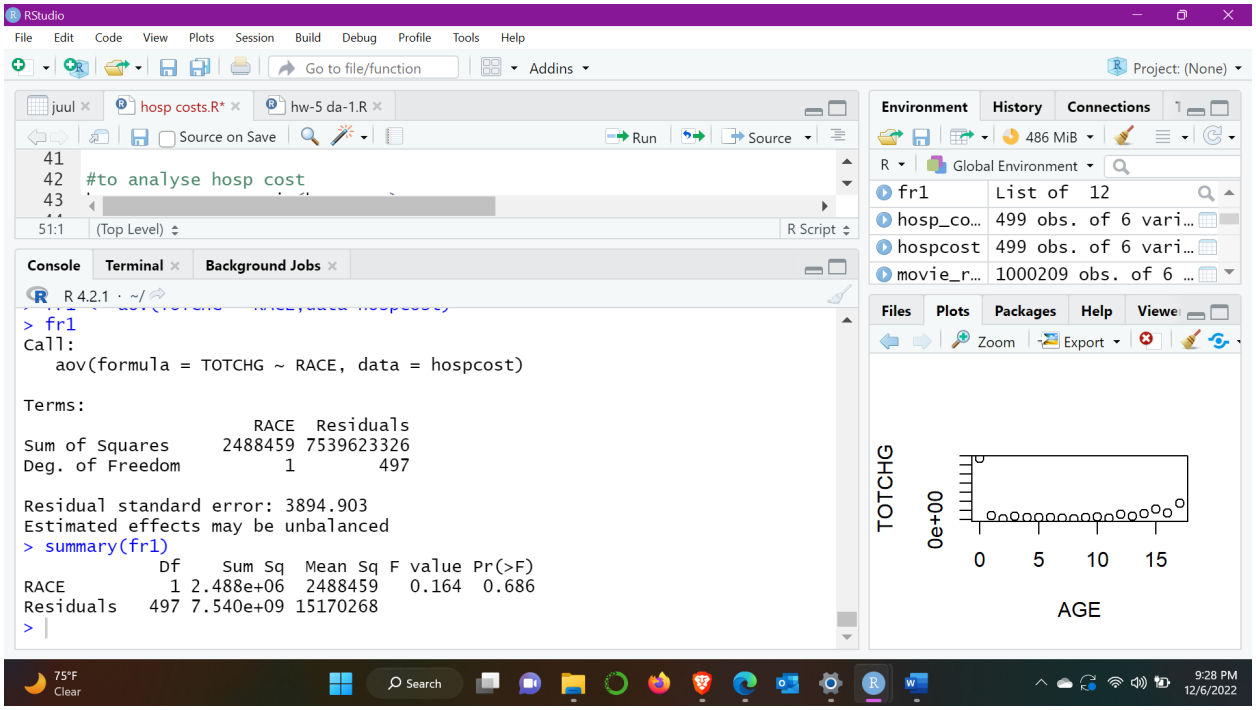

```
t <- table(hospcost$APRDRG)

d <- as.data.frame(t)

names(d)[1] = 'Diagnosis Group'

d

which.max(table(hospcost$APRDRG))

res <- aggregate(TOTCHG ~ APRDRG, data = hospcost, sum)

res

which.max(res$TOTCHG)

res[which.max(res$TOTCHG),]

t.test(res)
```



From t.test we see that p-value is fairly less than 5% hence it is statistically significant.

Step #4 Checking RACE wise hospital costs:

```
hosp_cost <- na.omit(hospcost)

hosp_cost$RACE <- as.factor(hospcost$RACE)

table(hospcost$RACE)

fr1 <- aov(TOTCHG ~ RACE,data=hospcost)

fr1

summary(fr1)
```



This shows that RACE doesn't contribute to expenditure. (Degree of freedom)

Step #5 Checking if age is a factor:

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

table(hospcost$FEMALE)

b <- lm(TOTCHG ~ AGE+FEMALE,data=hosp_cost)

summary(b)

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

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
```

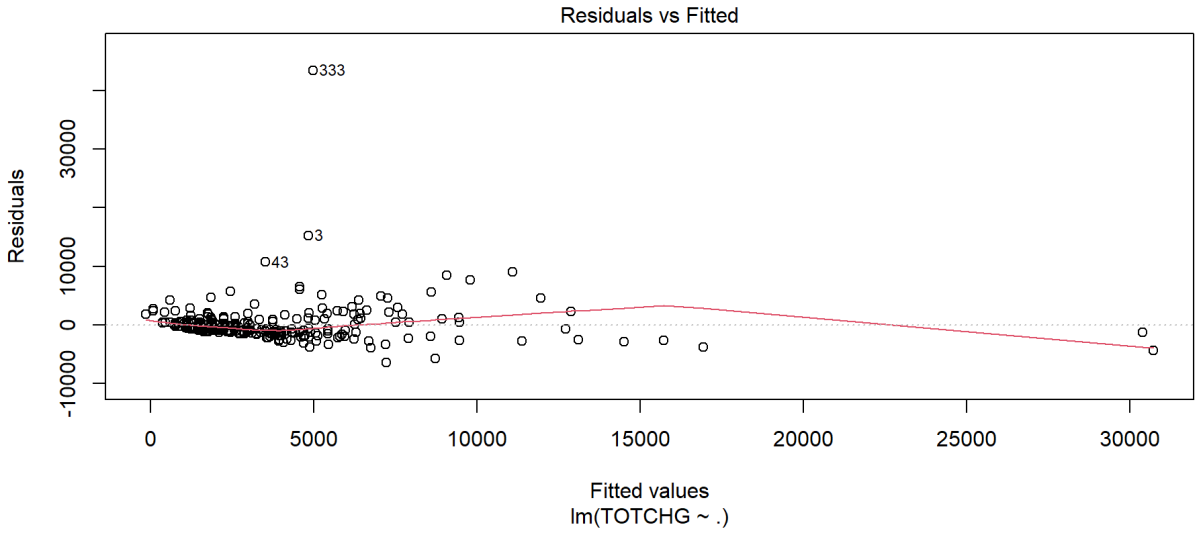
Step #6 Linear Modelling

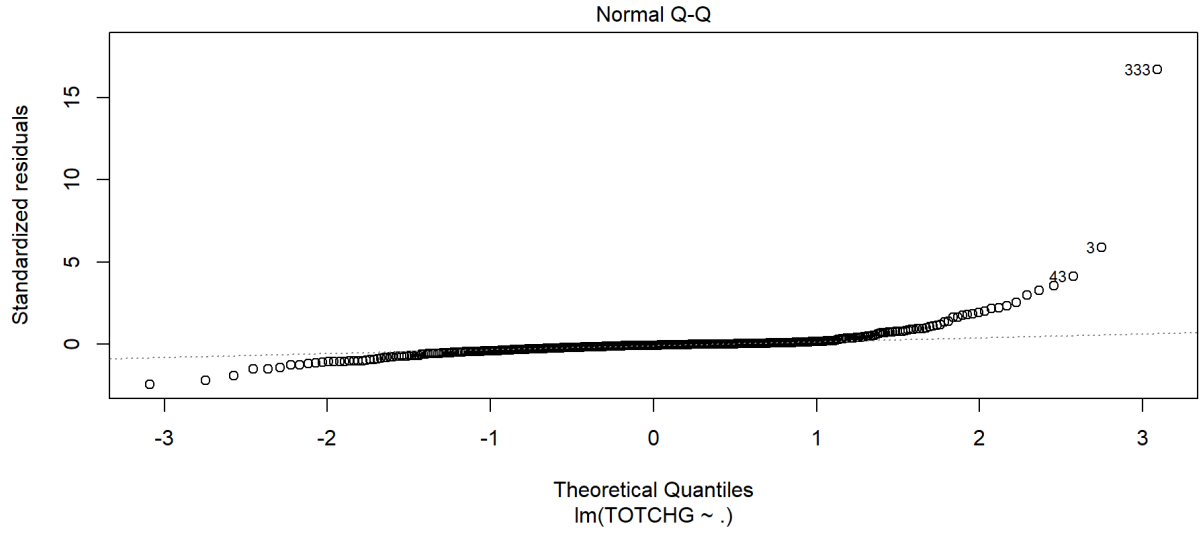
cost <- lm(TOTCHG ~ .,data=hosp_cost)

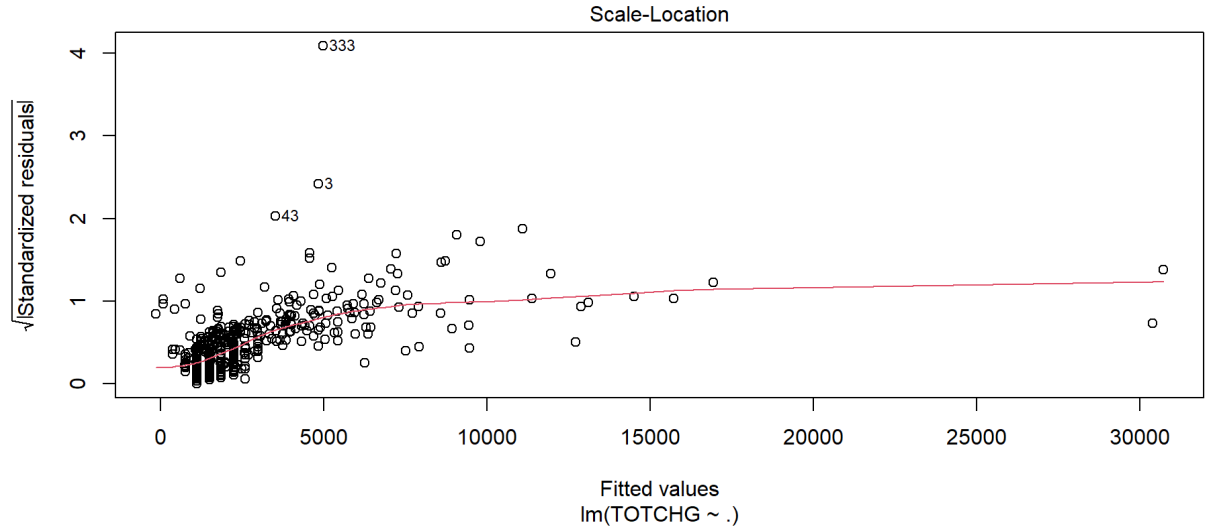
summary(cost)

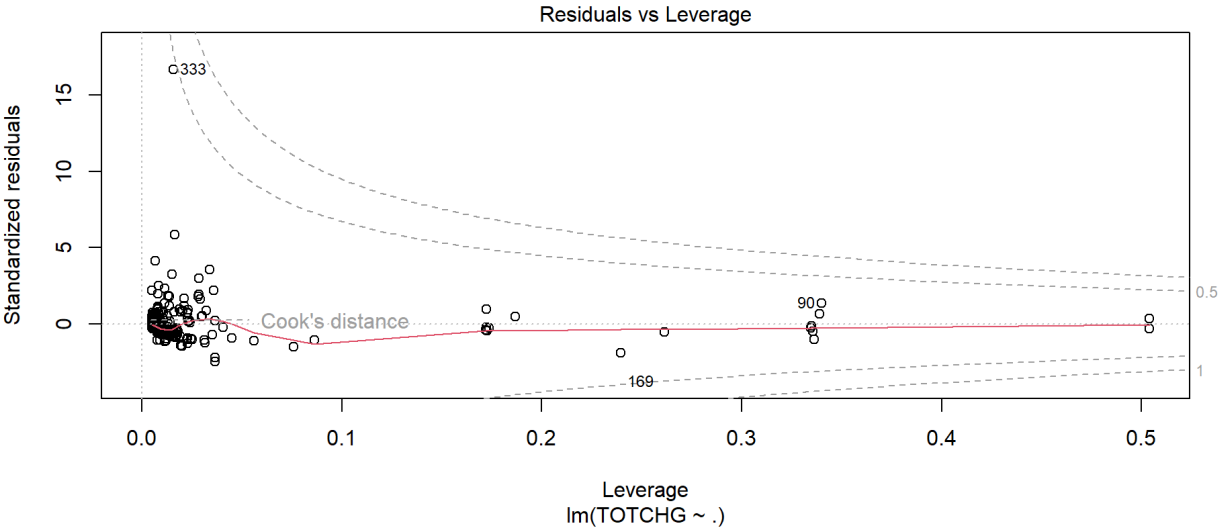
plot(cost)

Linear Model Plots









Call:

lm(formula = TOTCHG ~ ., data = hosp_cost)

Residuals:

Min	1Q	Median	3Q	Max

-6367 -691 -186 121 43412

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	5024.9610	440.1366	11.417	< 2e-16	***
AGE	133.2207	17.6662	7.541	2.29e-13	***
FEMALE	-392.5778	249.2981	-1.575	0.116	
LOS	742.9637	35.0464	21.199	< 2e-16	***
RACE2	458.2427	1085.2320	0.422	0.673	
RACE3	330.5184	2629.5121	0.126	0.900	
RACE4	-499.3818	1520.9293	-0.328	0.743	
RACE5	-1784.5776	1532.0048	-1.165	0.245	
RACE6	-594.2921	1859.1271	-0.320	0.749	
APRDRG	-7.8175	0.6881	-11.361	< 2e-16	***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2622 on 489 degrees of freedom

Multiple R-squared: 0.5544, Adjusted R-squared: 0.5462

F-statistic: 67.6 on 9 and 489 DF, p-value: < 2.2e-16

Results:

- There are more no. of patients with age between 0-1.
- P-value is less than 0.05, therefore it is statistically significant
- Fitted Model equation would be:

5024.9610+133.2207 *AGE-392.5778*FEMALE+742.9637*LOS+458.2427*RACE2+ 330.5184 *RACE3 - 499.3818RACE4-1784.5776*RACE 5-594.2921*RACE 6-7.8172*APR DRG.

- Most expenditure is done from age group 0-1 as they are more present in the hospitals.
- The residual standard-error shows out of 2622 there are 489 degrees of freedom which shows the linear model is fitted well.
- Additionally, there is a correlation between length of stay and hospital expense, leading to the conclusion that for every increase in length of stay of 1 the hospital expense rises by 742.9.
- For Race 1, we only have about 484 out of 500 values of data, which makes it less trustworthy. It can therefore be said that there was not enough data for a fair analysis to determine the relationship between hospital costs and patient race.

Conclusion:

1. Health care costs are dependent on age, length of stay and the diagnosis type.
2. Healthcare cost is the most for patients in the 0-1 yrs age group category
3. Length of Stay increases the hospital cost All Patient Refined Diagnosis Related Groups also affects healthcare costs
4. Race or gender doesn’t have that much impact on hospital cost.
5. These three factors will serve as the independent variables in our analysis so that we can easily determine if length of stay is influenced by age, gender, or race.
6. The P-value is fairly high for all the independent variables, indicating that there is no linear relationship between them. It follows that we cannot predict a patient's length of stay based on their age, gender, or race.
7. The frequency of hospital visits directly impacts how much is spent, therefore the higher the number of trips, the higher the cost. (As shown by the analysis above.)

References:

https://www.kaggle.com/

https://stackoverflow.com/

https://docs.posit.co/

https://www.geeksforgeeks.org/

https://principlestudies.org/essays/health-care-in-america/

https://www2.kenyon.edu/Depts/Math/hartlaub/Math116%20Spring2016/R.htm

https://stat.ethz.ch/R-manual/R-devel/library/stats/html/stat.anova.html