**ANALYZE HEALTHCARE COST AND UTILIZATION IN WISCONSIN HOSPITAL**

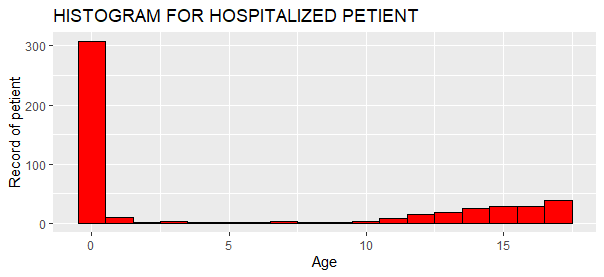
**Problem statement** –

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 the healthcare costs and their utilization.

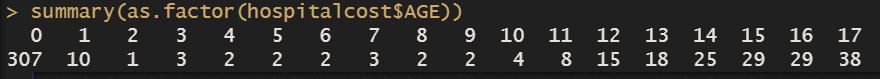
**The goals of the project**

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.**

**Results -**

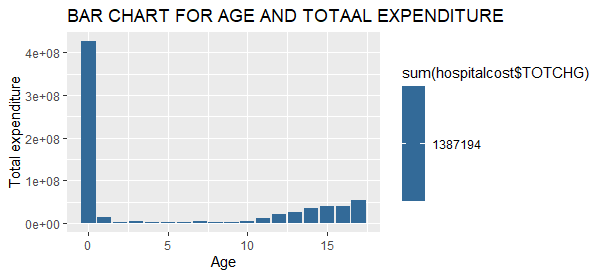


Numerical values of diagram

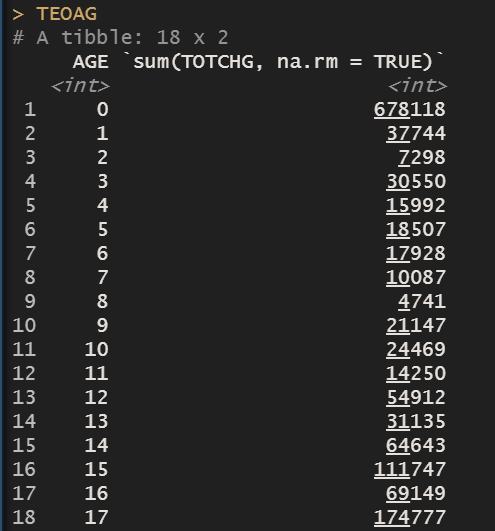


From the above diagram it is clearly evident that the 0 to 1 range age group has highest number of records available for hospitalization. The bar is surpassing the value of 300 which is given in the table below. To derive these results, the histogram is created by using qplot method and the results shown in table are calculated by summary method. In the table we can see that the 0-age group is having 307 numbers of records for hospitalization which is highest among all of other group.

The second query of the same problem statement the agency wants to find the which age group is having maximum expenditure, this problem is explain by the table and diagram given below



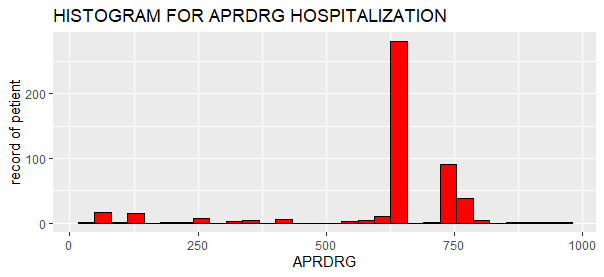
Numerical values of the above diagram



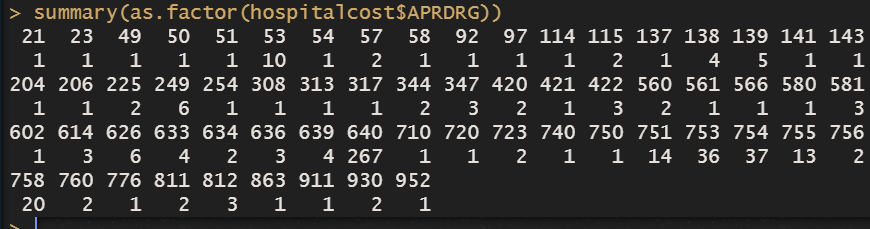
Here in the above table, we can see that the 0-age group has a maximum number of expenditure and this numerical is visible through diagram as well. After installing dplyr package the group\_by function is used to aggregate the values. The 0-age has 678116 total amounts of expenditure.

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.**

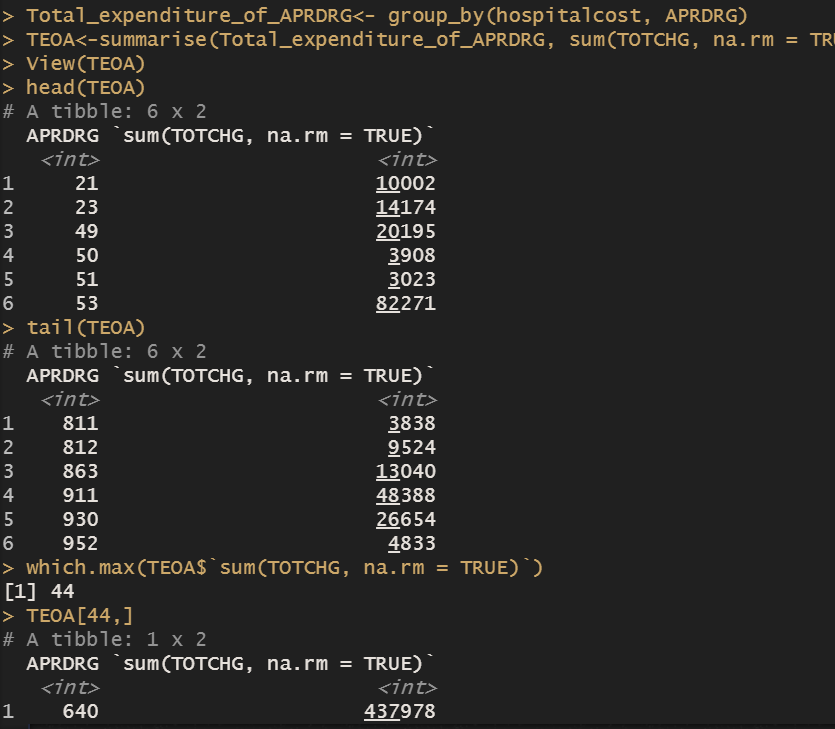
**Results -**



In the above diagram the APRDRG is mention on the x axis and the number of hospitalizations is recorded on the y axis. There are 64 APRDRG diagnosis related groups, most of them are having very low value which is shown is diagram but there is a one group between 500 to 750 which has maximum records compare to other. By examining the below table there is a 640 group which has record of 267 patients, this group has been hospitalized frequently.



Moving further to find a maximum expenditure of diagnosis related group the same methodology has been used, which is used for previous problem statements but in this case, there is large number of data recorded therefore I took a help of which.max function, this function help us to find the maximum value from the given data set. To understand this concept please follow the table given below. The following table also include some of the other functions like head() and tail() function, these functions shows us the top and bottom values of the data set. In the next line we can see by using which.max function we find that row which contains the maximum value for maximum expenditure. And at the end of table by putting the row number the value is arrive that is, 640 group has maximum expenditure of 437978.



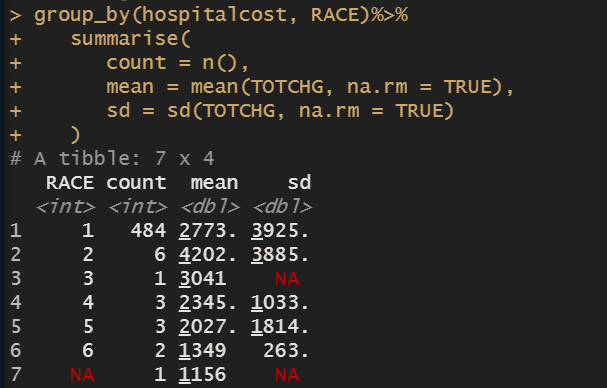
1. **To make sure that there is no malpractice, the agency needs to analyse if the race of the patient is related to the hospitalization costs.**

**Results –**

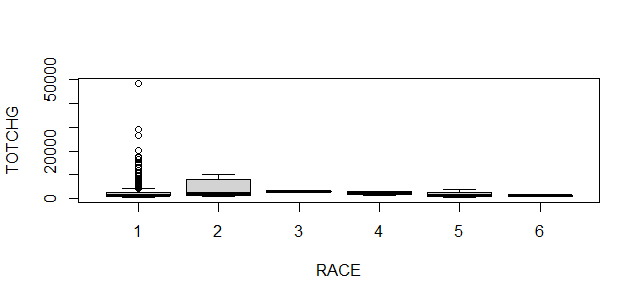
In the data set there are 7 race categories and number 1 race category has a maximum number of records. As per the above problem statement the agency wants to find is there any malpractice happening on basis of race, and to find the answer for this query I used ANOVA which is mainly used when we want to compare the difference of mean bitween the groups on the basis of perticular variable.

Here’s the racial category is compared, and with help of ANOVA I tried to find is there any difference of expenditure on basis of race. But before that let’s calculate the mean of the each race category and find the diffrences of their mean

To understand the clear view of numerical and graphical results please consider the table and diagram given below.



Here we can see there is seven racial categories and among them the first racial category contains the most nummber of count. In the given table the third and forth column shows the mean and standerd deviation of respective groups. The mean column clearly shows that there is very less diffrence of mean between the racial groups. The fourth column shows the standerd deviation of groups and there are third and seventh row is emty because there are only one value is recorded for these racial groups. The following diagram will give you the clear picture of these results.

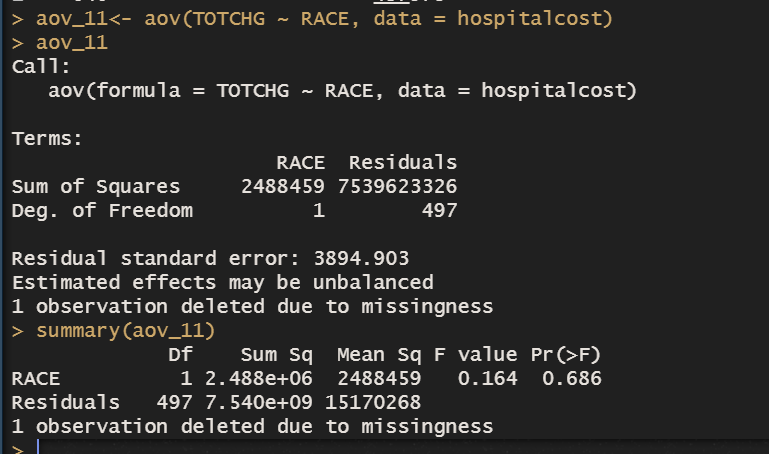


Here we can see the above diagram is showing the clear picture of the calculated result. As mention earlier there are seven racial category but here one of them is removed due to its missing value. In the diagram the race is plotted on the x axis and TOTCHG is mention on the y axis. Through the boxplot it is evident that there is very less diffrence of the mean between the groups. The dark line which is located uder the box represents the mean of the racial group with respect to TOTCHG.

To get conferm results let’s calculate the ANOVA but before that we need to set the null and alternate hypothesis. Here the hypothesis are,

H,0 = “ there is no significant diffrence of mean between racial groups”

H,1 = “ there is significant diffrence of mean bitween racial groups”



The above picture represents the detailed results of ANOVA, which is calculated after applying aov method. In the first row there is formula which takes TOTCHG as dependent factor and it compares with RACE which is categorical variable. The summary function gives us the summary of ANOVA results, here in first column we get digrees of freedom in term of race and residuals.

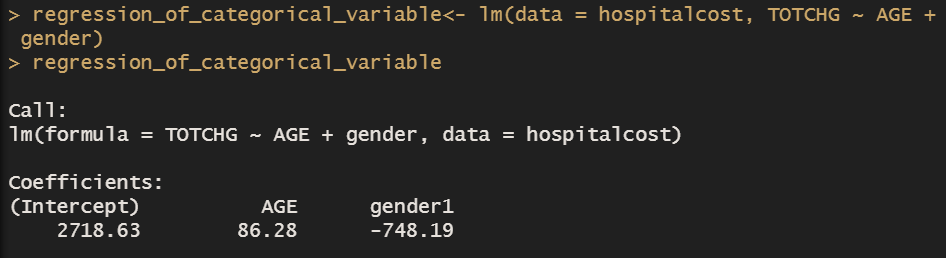
The RACE represents the values of df - 1 and the row below that shows the total number of observations - total number of digrees of freedom, apart from that there are few columns like sum of square and mean square. To get a mean of square we have to devide the df by sum sq. the mean of square is usefull when it comes to calculation of p-value. By deviding numerator and denominator of mean of sq we get the p-value. In the hypothesis testing when the p-value is greater then 0.05 then you have to accept the null hypothesis which says there is no variance or significance diffrence in variable.

In the above table we can see the p-values is 0.686 which is greater then 0.05. it means there is no significance diffrence of mean bitween the category of RACE. And here it conclude that there is no malpractice happening on basis of racial category.

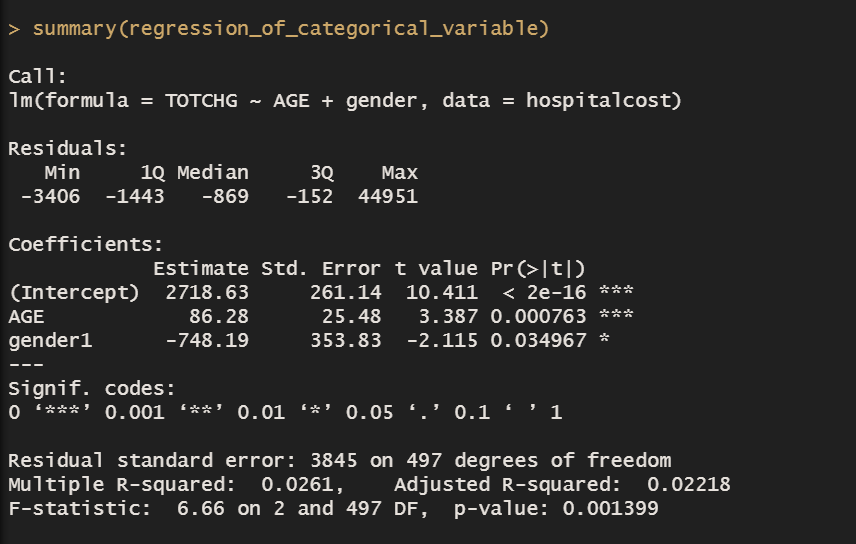
1. **To properly utilize the costs, the agency has to analyse the severity of the hospital costs by age and gender for proper allocation of resources.**

**Results-**

The severity of hospital cost as per the age and gender can be measure by the effect of age and gender on hospital cost. To solve the query we can use multiple linear regression model. The multiple linear regression helps us to find the effect of independent variable on dependent variable and tells us how the perticular variable is significant for the efficiency of model, here the TOTCHG is a dependent variable which is continuous in nature where as the age and gender is independent variable. In linear regression there are no as such prerequisite for independent variables data type and in this data set the age and gender are factorial variable. The age has 17 age groups and in gender column 0 represent as male and 1 represents to female.



The above picture shows us a compact overview of the result, in this picture the vector which is contain the linear regression model has gave us the only intercept term and beta coeffient, and the line below the code is showing the formula of lm model but this level of information is not sufficient therefore I used summary function to get a detailed information about this model.



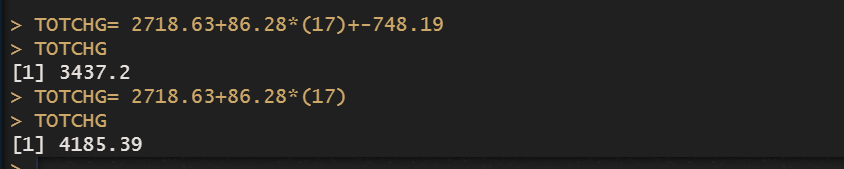
The first line discribe the formula of this model. In that formula we have TOTCHG as y variable which is dependent and age and gender are independent variable. The next line of result will explain us the residuals, the residual is nothing but an diffrence bitween the actual and predicted values of the regression model, here we can see the gap between the residual is quite high because there must be some high values has been found on above or below the regression line therefore the range of residual is high.

the next section of the model will gives the information about intercept and beta coefficients. The intecept is correspond to TOTCHG, it explains that when all beta coefficient is zero then the model will have this initial value in form of intercept. The next row from this column will tell you beta slop or coefficient, it means when one unit increase in AGE the TOTCHG will increase by 86.28 and in case of gender1 it will effect the TOTCHG in dicreasing manner due to its negative relation. Here the gender is dummy veriable therefore the other category is removed and female is considered as dummy veriable.

The another task which is done by lm camand, it provides level of significance which tell’s wether the the independent variable is effective for dependent variable or not. There are many ways to find it, like checking t-distribution which we get after deviding std error with all its respective estimator. If the t-distribution comes under the interval of 2 to -2 then it is clear that the independent veriable is not significantly different then zero. And from above result we can see that the t-distribution has a higher value the 2 to -2 interval. From other way we can decide the significance level, is by looking at p-value, if the p-value is lesser then 0.05 then we can say that the independent variable is significantly different then zero and in this case we have both the values of independent variables are significant in nature because it has the p-value lesser then 0.05.

So from above interpretation it is clear that the both veriables are significant in nature and they are affecting the dependent variable.

Thruogh the problem statement the agency wants to find, how much the AGE and gender is affecting the TOTCHG (hospital cost) and by interpreting above result we get beta coefficient which tell us the rate of change of dependent veriable by given independent verable. The following example will explain this concept in brief manner.



In the above examples to calculate the TOTCHG the formula is used as,

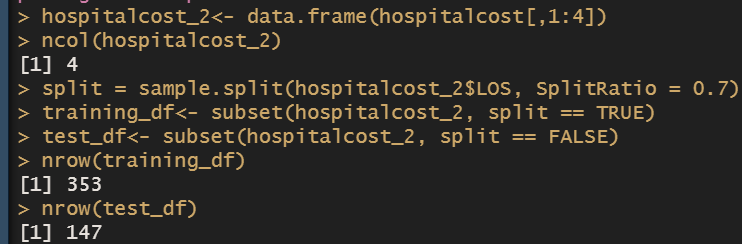
TOTCHG= intercept + beta1\_coefficient\*(AGE) + beta2\_coefficient\*(GENDER)

And here we get the vaules, the first example has a age of 17 and takes the gender1 which is female, the result of TOTCHG is chenging with respect ot chenge in coefficient of AGE and GENDER1. The answeer for the first equation is different because here we take female as 1 and male as zero.

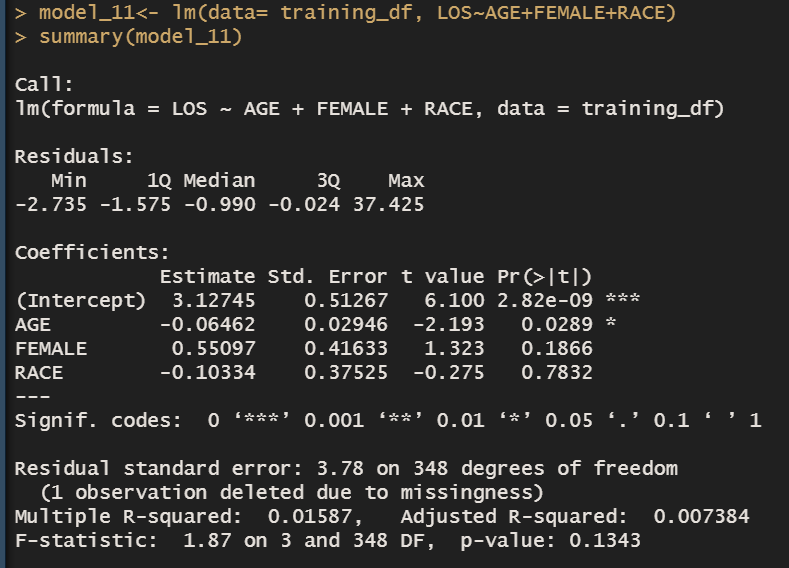
1. **Since the length of stay is the crucial factor for inpatients, the agency wants to find if the length of stay can be predicted from age, gender, and race.**

**Result –**

To solve this query I used multiple linear regression model and try to predict the LOS on basis of AGE, GENDER, RACE. So here the los is dependent variable and age, gender and race is independent variable. But before making calculation lets prepare a data for the model. The following image will provide the codes from which I prepare the data for further model.



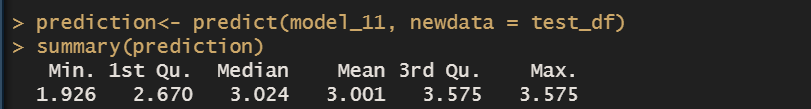
The above iamge is giving us all the information like how the data is created for the model and its number of columns, the next few rows is about methodology of spliting data into train and test data set with 70 30 split retio, and therefore we have 353 rows in training dataset and remaining rows are included in test data set. The following model is created on the basis of training data set. By putting the values into the model here what comes out in form of result.



The most of the concepts are explained in previous query like the first line describes about the formula and the values which is put into the formula, as mention earlier the y variable is LOS and the variables next to it are all independent variable. Moving further we have the intercept and beta coefficient column which tells the rate of change of dependent variable due to independent variable but her the most important thing we have to notice in this model is the level of significance. And from p-value we can find it. Here from above results, it is clearly evident that the only the age has higher level of significance and effect on the TOTCHG because the p-value is lesser then 0.05.

Which means the age is not significantly different from zero and has effective relation with the LOS.

But as mention in the problem statement the agency wants to find the prediction values of LOS on basis of age, gender and race. So, the following result will give the answer for this query. The predicted value of LOS has been calculated by taking the test data set and the summary function comes with insides of data set.

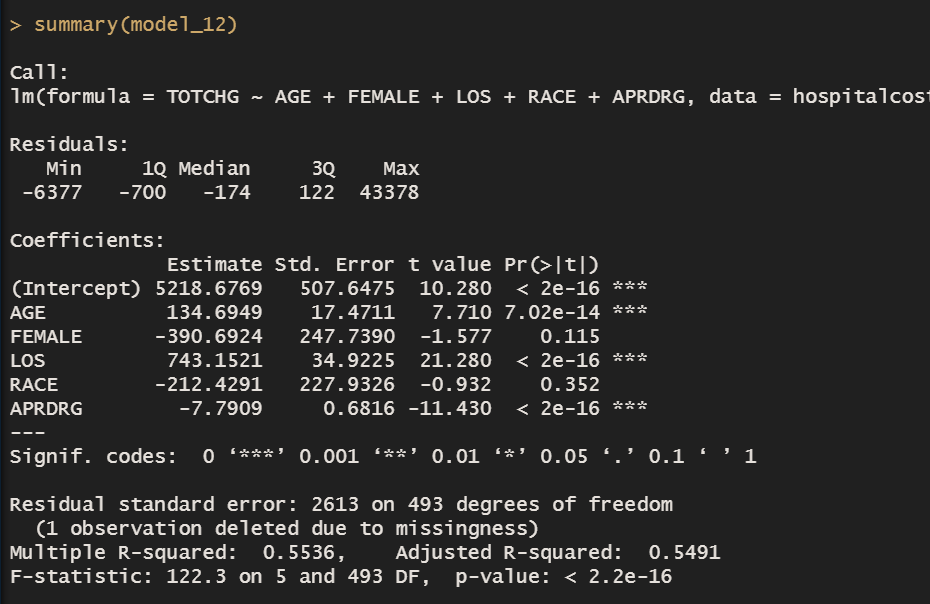


The result shown in the above picture has a predicted value of the LOS and summary function summaries the entire values and gives the mean of the los and minimum maximum records from the test data set observation. Here we can see that the value for length of stay is depict in range of 1 to 3 days on basis of age, race and gender. The comparison of actual and predicted values is given in index section of this project. Instead of test data set, the training data set is used for comparison of predicted and actual values of LOS.

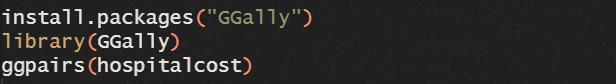
1. **To perform a complete analysis, the agency wants to find the variable that mainly affects the hospital costs.**

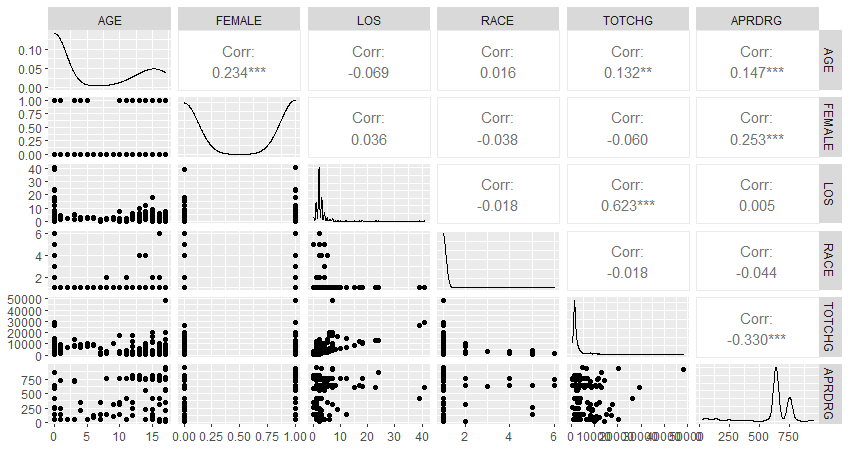
**Result –**

There are mainly five variables apart from los in data set to find the effectiveness of those variable the correlation plot can be useful but it won’t gives information in numerical structure. The multiple linear regression has a capability to provide the information about effectiveness of the variable for dependent variable. after putting the formula, we have results of a LM model. Here the results are given bellow.



The first line shows the formula and the line next to it describes the information about residuals, (the difference between the actual and predicted values of dependent variable) but in this case the agency wants to find the effectiveness of particular variable with respect to dependent variable and for that we can look at the values of intercept and beta coefficient of independent variable. the above picture gives the information through the intercept column that there are two variable which has a positive impact on the hospital cost and among them the LOS (length of stay) has higher influence on the hospital cost but another question arise, is it significant for this model? and to answer this let’s look at the p-values of this model. In the p-value column the p-value for the LOS is very less then the 0.05 which means the variable is significant in nature and it is effective for this model.

There is another way as well like checking correlation through pair plot these are all commands has been used to create pair plot.



The above plot tells the results of correlation in numerical format and on the other hand the relation of the variable is plotted with the help of scatter plot. From the above results the LOS has a higher correlation with TOTCHG and rest of other variable has a negative values or weak correlation

In correlation when the value is near to one then it has a good correlation whereas when values is moving towards zero it indicates that the variable has weak correlation. And the negative values show the decreasing or disperse relation between the variable.

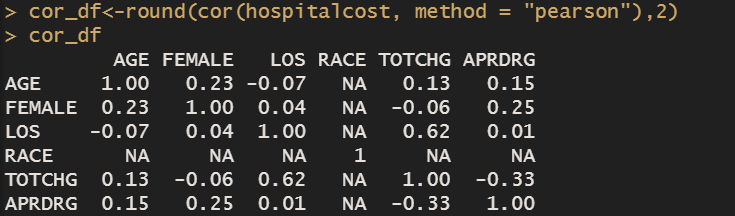
By considering the above results the LOS and TOTCHG has correlation of 0.623 which means the relation between these two variables is positive in nature and the correlation between them is highest among all other correlation so it is clearly visible that the TOTCHG has been highly affected by the length of stay.

**Summary –**

* the 0-age group is having 307 numbers of records for hospitalization which is highest among all of respective groups. And the 0-age has 678116 total amounts of expenditure which is highest compare to all other age group.
* The 640 number of groups which has record of 267 patients, this group has been hospitalized frequently, and it has maximum expenditure of 437978.
* There is no malpractice happening on basis of racial category.
* The result of TOTCHG is chenging with respect ot chenge in coefficient of AGE and GENDER1.
* The length of stay is depicted in range of 1 to 3 days on basis of age, race and gender.
* The length stay has positive influence on TOTCHG (hospital cost) as compares to independent variable.

Index –

The correlation of hospital cost data set



The predicted values of LOS by using the training data set

