Project 1:
Healthcare cost analysis

Abhilash V

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

Expected Outcome:

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.
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 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 the hospital costs.

Workflow Description

I have used R-language to work on the project and analyze the data.

Data Set has been downloaded from the SimpliLearn

Function Explanations

Summary() – To give a summary of each attribute in the Data set. It includes:

Min. value, 1st Qu. Value, Median value, Mean value, 3rd quartile value, Max value

Hist()- to plot the histograph

Factor()- used to create factors (vector; has a sorted order)

Aggregate()- used to create a data frame and displaying the result on the basis of given formula (FUN)

Annova- Package function is being used for Analysis of Variance

Linear Regression- Package function is being used to find the relationship between factors

Analysis

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.

Based on the output we can see that Age wise Hospital Visit and Expenses

Maximum Hospital Visit – 0-1 yrs age group : 306

Maximum Expenditure – 0-1 yrs age group : 676962

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.

Based on the output we can see the list of Expenditure based on the Diagnosis and treatment.

TOTCHG - Hospital discharge costs

APRDRG - All Patient Refined Diagnosis Related Groups

640 (All patient refined Diagnosis Related Gp) has the maximum expenditure - 436822

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.

Annova test is being used to analyze the Race wise cost occurred

The Residual Value (deviation of the observed value) is very high specifying that there is no relation between the race of patient and the hospital cost.

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

This will affect the results of ANOVA as well, since the number of observations is very much skewed.

Hence we can conclude that there is no race wise cost bias in the observed data.

4) 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.

Linear Regression Model has been used

Age is a very important factor in the hospital costs as seen by the significance levels and p-values.

The gender also seems to have an impact.

There is an equal number of male and female patients.

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

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.

Linear Regression Model has been used

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

The p-value high which signifies that there is no linear relationship between the given variables.

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

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

Linear Regression model has been used

Based on the output we can see that the Age and Length of stay affects the total Hospital cost.

Cost is directly proportional to the Length i.e. higher the Length of stay of patients will result to higher hospital cost.

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