MISSION HOSPITAL CASE STUDY ANALYSIS



Name of Student (1): Souvik Das

Name of Student (2): Varsha Satpathy

Name of Student (3): Sam John

Name of Student (4): Nikhil Shah

1. Write Case Summary of Mission Hospital

Mission Hospital is a Super Speciality Hospital in Durgapur, West Bengal specialized in Cardiovascular surgeries and they are trying to decentralize the entire management system by initiating “Packaged Pricing” for the patients coming from different parts of the country and neighbouring countries. The hospital is considering optimizing its model to increase customers' confidence as well as make its pricing policy more accurate so that hospital won’t incur losses.

The flat prices for various treatments are becoming common in Indian hospitals, mostly owing to health tourism and government institutions. Dr. Satyajit Bose, President of Mission Hospital, is considering implementing a flat-rate approach to give a patient a fixed price for their therapy on admission. This will mean that based on the patient's clinical and non-clinical details available at the time of admission, the package price should be determined by the hospital. Flat fees (or packaged pricing) may cause hospital loss if the care expenses of patients are not correctly measured. It is crucial for the Hospital to determine the major factors driving the actual cost of care to quote the patient for an approximate rate of treatment.

1. Identify the problems in the dataset and suggest the measure to clean it?

There are dummy variables created in the data but some of them are not created or made, hence need to create the remaining dummy variables.

Missing / NA values: As it is medical data, it is a good practice to remove the NA/missing data instead of putting some random values at that place. Mean, KNN or any kind of imputation is very risky for a medical health data so it is best not to impute but to remove the rows. But in this situation if we remove all missing and NA values we left with only 25% of data. Hence, omitting these values also won’t be a feasible solution.

(For now, we have kept NA values as it is as they are not a part of our regression model in the given situation.)

1. Develop a Simple Linear Regression to check if there is association between Total Cost and Body Weight?

**After building the Simple Linear Regression model, we found that the association between the Total Cost and Body weight is very low i.e 11.78% thus not a very good model for prediction but the independent variable is significant here.**

The predicted model is given by,

**Y = 129397.7 + 1846.9X**

*Below is the code and output for the same:*

lm(formula = TOTAL\_COST\_TO\_HOSPITAL ~ BODY\_WEIGHT, data = Mission\_Hospital)

Residuals:

Min 1Q Median 3Q Max

-191713 -64862 -25233 24508 647139

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 129397.7 13961.1 9.268 < 2e-16 \*\*\*

BODY\_WEIGHT 1846.9 316.9 5.829 1.74e-08 \*\*\*

---

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

Residual standard error: 115100 on 246 degrees of freedom

Multiple R-squared: 0.1214, Adjusted R-squared: 0.1178

F-statistic: 33.98 on 1 and 246 DF, p-value: 1.743e-08

1. Find the correlation between variable "Age", "Body Weight", "Body Height", "Total Length of Stay", "Length of Stay ICU", "Cost of Implant", "Total Cost to Hospital".

From the below graph we can see that the

Correlation between Age and Body Weight is 0.85, highly positive correlation

Correlation between Age and Body Height is 0.72, highly positive correlation

Correlation between Body Height and Body Weight is 0.85, highly positive correlation

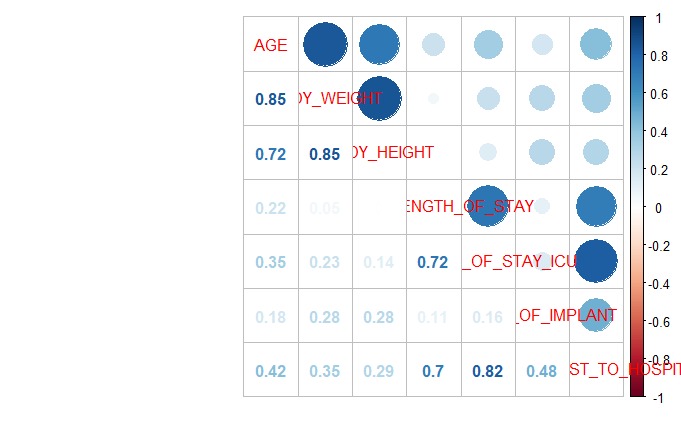
Correlation between Age and Length of Stay is 0.22, low positive correlation

Correlation between Age and Length of Stay in ICU is 0.35, low positive correlation

Correlation between Age and Cost of Implant is 0.18, low positive correlation

Correlation between Age and Cost to Hospital is 0.42, low positive correlation

And so, on



1. Develop a forward Multiple Linear Regression using the relevant variables given in question 4, and identify statistically significant predictors that mission hospital can use to find Treatment Cost? Also do the heteroscedasticity analysis and write the report?

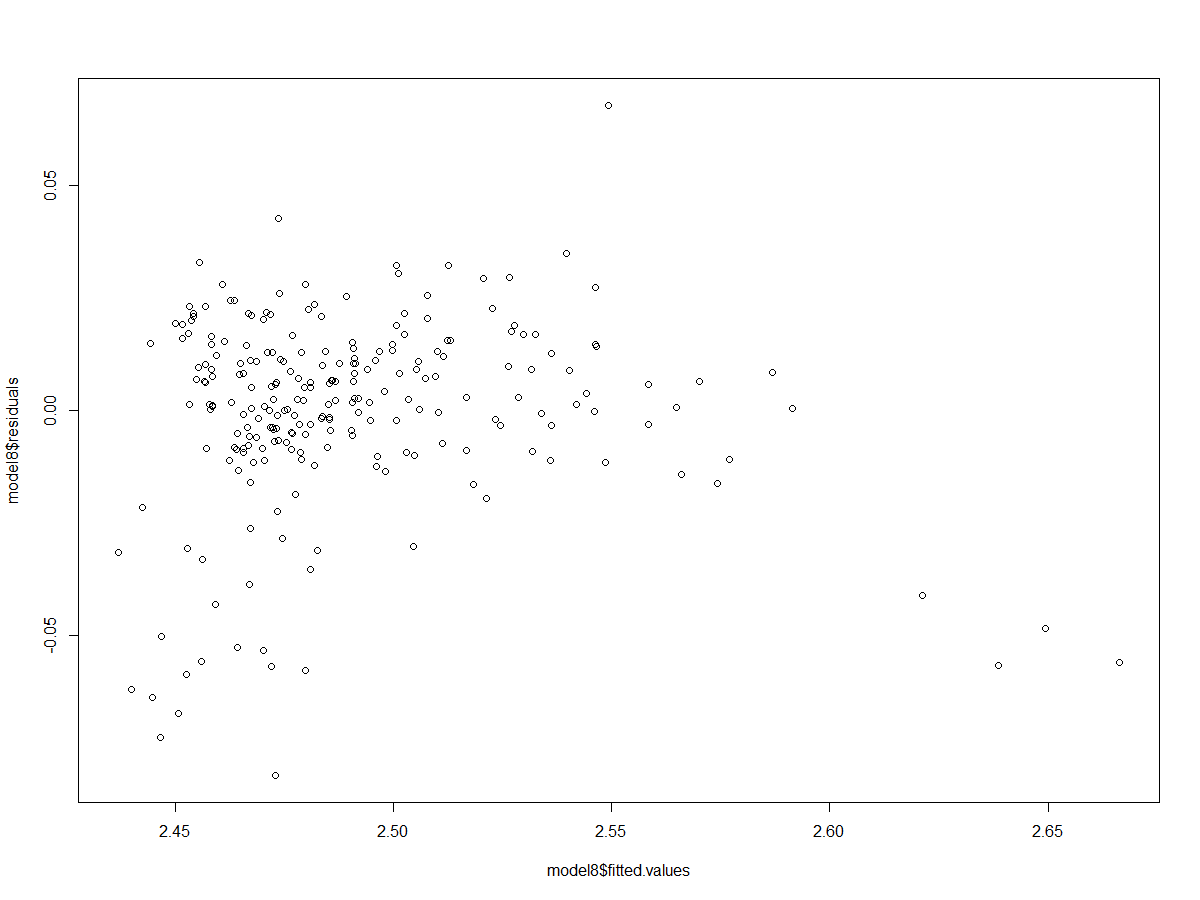
Initially, the response or dependent variable “TOTAL\_COST\_TO\_HOSPITAL” was not normally distributed and positively skewed so we took log transformation to make the data normally distributed. Below is the given graph



Now building the models keeping TOTAL\_COST\_TO\_HOSPITAL as response variable and doing forward linear regression

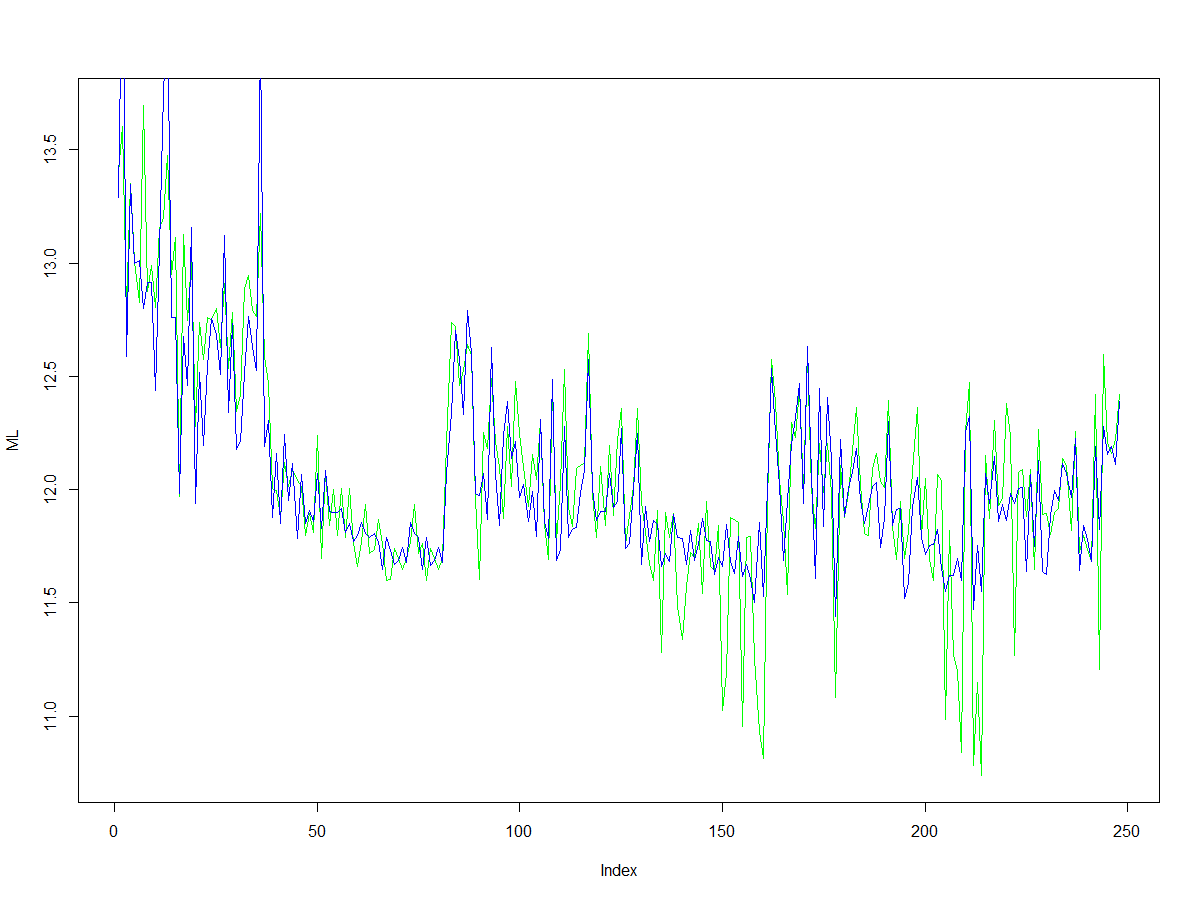
|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **IDV** | **Significance** | **Adjusted R square** |
| Model1 | AGE | \*\*\* | 0.1869 |
| Model2 | AGE | \*\*\* | 0.1845 |
| BODY\_WEIGHT | Not Significant |
| Model3 | AGE | \*\*\* | 0.1847 |
| BODY\_HEIGHT | Not Significant |
| Model4 | AGE | \*\*\* | 0.5537 |
| TOTAL\_LENGTH\_OF\_STAY | \*\*\* |
| Model5 | AGE | \*\*\* | 0.621 |
| TOTAL\_LENGTH\_OF\_STAY | \*\*\* |
| LENGTH\_OF\_STAY\_ICU | \*\*\* |
| Model6 | AGE | \*\*\* | 0.7268 |
| TOTAL\_LENGTH\_OF\_STAY | \*\*\* |
| LENGTH\_OF\_STAY\_ICU | \*\*\* |
| COST\_OF\_IMPLANT | \*\*\* |
| Model7 | BODY\_WEIGHT | \*\*\* | 0.7346 |
| TOTAL\_LENGTH\_OF\_STAY | \*\*\* |
| LENGTH\_OF\_STAY\_ICU | \*\*\* |
| COST\_OF\_IMPLANT | \*\*\* |
| Model8 | BODY\_HEIGHT | \*\*\* | 0.7357 |
| TOTAL\_LENGTH\_OF\_STAY | \*\*\* |
| LENGTH\_OF\_STAY\_ICU | \*\*\* |
| COST\_OF\_IMPLANT | \*\*\* |

After performing forward linear regression, we got Model 8 as the best fitted model. And there is **no heteroscedasticity and multicollinearity** in the model, the graphs are shown below.



As we can see from the above graph there is no funnel pattern in the model8, hence **no heteroscedasticity** in the model.

Below is the graph of **actual vs predicted** values.



Some examples of actual vs predicted values

|  |  |  |
| --- | --- | --- |
| **Observations No (HEAD)** | **Actual Value** | **Predicted Value via Model8** |
| 1 | 13.40044 | 13.28793 |
| 2 | 13.60371 | 14.38844 |
| 3 | 12.80004 | 12.58583 |
| 4 | 13.35346 | 13.34838 |
| 5 | 13.00555 | 12.99764 |
| 6 | 12.82761 | 13.01196 |

The model8 is giving above 70% accuracy.