GOA INSTITUTE OF MANAGEMENT
PGDM (Big Data Analytics)
Predictive and Prescriptive Analytics
2020-21

MISSION HOSPITAL CASE STUDY ANALYSIS

Name of Student (1): SANDIP GANGULY_B2020045

Name of Student (2): SARTHAK AGARWAL_B2020046

Name of Student (3): SARTHAK DATTA_B2020047

1. Write Case Summary of Mission Hospital

Mission Hospital case study highlights the problems faced by the hospitals in adopting the package system for the incoming patients. The underlying problem is that the hospitals are unable to predict the accurate cost of the package resulting in the hospital to lose money in case a patient overstays in the hospital and require care for a few extra days than expected/predicted. On the other hand, if the Hospital doesn't adopt to the package system then payment to be received after the treatment is a hassle for them as the clients are unwilling to pay for the amount charged as they believe there are some tests which are conducted unnecessarily to spike up the bill. In lieu to solve these problems the Hospital want to incorporate a more accurate prediction model for calculating the total cost incurred by the hospital so as not to run into losses.

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

1. NULL VALUES(Numerical) in the Dataset (Exhibit 2.1)
   a. BP_HIGH: 23 i.e. 9.27%
   b. BP_LOW: 23 i.e. 9.27
   c. HB: 2 i.e. 0.81%
   d. UREA: 13 i.e. 5.24%
   e. CREATININE: 33 i.e. 13.30%

Solution: Imputing the missing values with the ideal values of the variables (Exhibit 2.2)

2. AGE (years) variable has values which are in decimal (for infants/below 1 years) (Exhibit 2.3)

Solution: Correcting it with approximation **if Age≤0.5 = 0.5** or **if Age>0.5 = 1**

3. MARITAL STATUS of the person with AGE = 0.83 is set to Married (Exhibit 2.4)

Solution: changing the marital status of this record with **UNMARRIED** as this record is of an infant.

4. KEY COMPLAINT CODE has missing values (Exhibit 2.5)

Solution: imputing the missing values in this column either with **Unavailable/Under-Diagnosis**

5. In PAST MEDICAL HISTORY column code given to the past medical history of the patient are **Diabetes1**, **Diabetes2**, **Hypertension1**, **hypertension2, hypertension3**, **Other** ideally but after studying the column we come to know that there is one more code is there in the data i.e. **hypertension1**(which might be a data entry error).(Exhibit 2.6)

Solution: We impute the **hyptertension1** with **Hypertension1** (As may be a manual data entry error)

6. PAST MEDICAL HISTORY has missing values (Exhibit 2.7)

Solution: imputing the missing values in this column with **Unavailable.**

7. The TOTAL COST TO HOSPITAL column is not normally distributed (required to be normally distributed as per linear regression assumption) and has outlier as well. (Exhibit 2.8)

Solution: By Scaling this column it will get normally distributed (Exhibit 2.9)

8. When we plot a graph between total amount billed to the patient and concession given, we see three outliers present over there first, where the bill amount is not much high and, but the maximum concession is given even after going through an implant. Second, where the billing amount is highest, and the concession is second high. Third, here the bill amount is very less and given a very high concession though there is no implant as well (Exhibit 2.10)
9. When we plot a graph between Total cost to hospital and amount billed to the patient there is a presence of outliers (Exhibit 2.11)
10. When we plot a graph between Total cost to hospital and amount billed to the patient there is a presence of outliers (Exhibit 2.12)
11. Person with Age 0.83 has a weight of 78 Kg and Height of 173 which ideally is not possible. (Exhibit 2.13)

Solution: Imputing the weight and Height of this infant with the ideal weight an infant of 8 month have.

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

We took the required Variables i.e. **Total Cost to the Hospital** and **Body Weight** from the master Dataset **mission_hospital** to perform Simple linear Regression,

After performing Simple Linear Regression on the data, we found the observations:

a. The p-value of Body Weight comes out to be 1.74e-08 which is highly significant
b. Adjusted R^2 value comes out to be 0.1178, so the model taken is not good
c. We have one independent variable, the multiplied R^2 value comes out to be 0.1214.

Since the Adjusted R^2 value is less, we can say the model is not good. Given from this value we can say that Body Weight can handle only 11.78% variation in the Total Cost, so the association between them is very less.

(Exhibit 3.1)

4. 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 Correlation Plot we can have the below observations:

a. Age, Body Weight and Body are highly correlated to each other.
b. Total length of stay is the summation of Length of stay in ICU and the Length of stay in ward. So, Length of stay in ICU and Total length of stay is highly correlated to each other.
c. We can see that Total cost to Hospital is correlated to Length of stay in ICU, in turn, which makes it correlated to Total length of stay. Also Cost of Implant is slightly correlated to Total cost to Hospital. The other factors, Age, Body Weight and Body Height are very less correlated to Total cost to Hospital.
d. The Body_Weight and Body_Height has a High Multi Collinearity.

5. 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?

Developing the Forward Regression Model using the following variables,

Independent Variables: "Age", "Body Weight", "Body Height", "Total Length of Stay", "Length of Stay ICU", "Cost of Implant"
Dependent Variables: "Total Cost to Hospital".
We took the modified columns (mentioned above) from data frame after cleaning,

After performing Multiple Linear Regression on the data, we found the observations:

a. The p-value of Total Length of Stay, Cost of Implant and Length of stay at ICU comes out to be **<0.001** which shows these variables are highly significant wrt the regression model.
b. Adjusted R^2 value comes out to be **0.8386** so the model considered to be good with a fit of **83.86%** but as the dependent variable is not normally distributed so even after getting the higher R^2 value our model will suffer from the heteroscedasticity which means the predicted value for higher and lowest Total Cost to Hospital will have a lot of variance. (Exhibit 5.1)
   ***Different Independent variable wise Model interpretation/variation** **(Exhibit 5.10)**

Multicollinearity Removal:
Since there is a presence of Multicollinearity between Body Height and Body Weight, we can drop these two columns from our Model as their significance to the model is not much i.e. not contributing much to the model. Though by doing this we would be able to remove the Multicollinearity from our model. (Exhibit 5.2)
Again, after performing Multiple Linear Regression on the data excluding the Body Weight and Body Height, we found the observations:

a. The p-value of Age, Total Length of Stay, Cost of Implant and Length of stay at ICU comes out to be **<0.001** which shows these variables are highly significant wrt the regression model.
b. Adjusted R^2 value comes out to be **0.8386** so the model considered to be good with a fit of **83.65%** but as the dependent variable is not normally distributed so even after getting the higher R^2 value our model will suffer from the heteroscedasticity which means the predicted value for higher and lowest Total Cost to Hospital will have a lot of variance.

After Removing the Multi Collinearity we go for Heteroscedasticity Analysis –

1. Firstly, we conclude that there is a presence of heteroscedasticity as the Dependent Variable is not Normally distributed since resulting in the situation of heteroscedasticity. (Exhibit 5.3)
2. After plotting the graph between the fitted values and the original values of the dataset we saw presence of a funnel type formation thus concluding the presence of Heteroscedasticity. (Exhibit 5.4)
3. We perform Breusch Pagan Test (using Chi Square for the test of Variance between the fitted values and original values and the p-value) to check the Presence of Heteroscedasticity.
   (Higher the value of chi square more the variance or if the p-value <0.05, it means presence of Heteroscedasticity). (Exhibit 5.5)
   Here,
   Total Chi Square Value = 571.1257518 and p- value = 2.746874e-122

New Model Solving problem of Heteroscedasticity,

Now the solution of this Heteroscedasticity problem is by Normalizing the Dependent Variable i.e. Total Cost to Hospital. Normalizing by performing Log Transformation over Total Cost to Hospital.
So, performing the Multiple Linear Regression over the following variables,
Independent Variables: "Age", "Body Weight", "Body Height", "Total Length of Stay", "Length of Stay ICU", "Cost of Implant".
Dependent Variable: " Ln. Total.Cost.”.
We took the modified columns (mentioned above) from data frame after cleaning,
After performing Multiple Linear Regression on the training data, we found the observations:

a. The p-value of Intercept, Total Length of Stay, Cost of Implant and Length of stay at ICU comes out to be **<0.001** which shows these variables are highly significant wrt the regression model.
b. Adjusted R^2 value comes out to be **0.7555** so the model considered to be good with a fit of **75.55%.** (Exhibit 5.6).

Multicollinearity Removal from this model:
Since there is a presence of Multicollinearity between Body Height and Body Weight, we can drop these two columns from our Model as their significance to the model is not much i.e. not contributing much to the model. Though by doing this we would be able to remove the Multicollinearity from our model. (Exhibit 5.7)
Again, after performing Multiple Linear Regression on the data excluding the Body Weight and Body Height, we found the observations:

a. The p-value of Age, Total Length of Stay, Cost of Implant and Length of stay at ICU comes out to be **<0.001** which shows these variables are highly significant wrt the regression model.
b. Adjusted R^2 value comes out to be **0.7469** so the model considered to be good with a fit of **74.69%** but as the dependent variable is not normally distributed so even after getting the higher R^2 value our model will suffer from the heteroscedasticity which means the predicted value for higher and lowest Total Cost to Hospital will have a lot of variance.

Heteroscedasticity Analysis on this new model with Normalized Dependent Variable and no Multicollinear Variables–

1. We perform Breusch Pagan Test (using Chi Square for the test of Variance between the fitted values and original values and the p-value) to check the Presence of Heteroscedasticity.
   (Higher the value of chi square more the variance or if the p-value <0.05, it means presence of Heteroscedasticity). (Exhibit 5.8)
   Here,
   Total Chi Square Value = 30.103197728 and p- value = 4.663285e-06
   The Variance between the fitted value and the predicted value goes down after normalizing the dependent variable thus reducing the Heteroscedasticity.
2. After plotting the graph between the fitted values and the Residuals of the dataset we saw absence of funnel type formation thus concluding the absence/reduction of Heteroscedasticity. (Exhibit 5.9)

GOA INSTITUTE OF MANAGEMENT
PGDM (Big Data Analytics)
Predictive and Prescriptive Analytics
2020-21

Exhibit 2.1

```
sapply(mission_EDA, function(x) (sum(is.na(x))/length(x)*100))
                      SL.                                AGE                             GENDER
                0.0000000                          0.0000000                          0.0000000
           MARITAL.STATUS              KEY.COMPLAINTS..CODE                        BODY_WEIGHT
                0.0000000                          0.0000000                          0.0000000
              BODY_HEIGHT                           HR.PULSE                           BP..HIGH
                0.0000000                          0.0000000                          9.2741935
                   BP.LOW                                 RR          PAST.MEDICAL.HISTORY.CODE
                9.2741935                          0.0000000                          0.0000000
                       HB                               UREA                         CREATININE
                0.8064516                          5.2419355                         13.3064516
          MODE.OF.ARRIVAL          STATE.AT.THE.TIME.OF.ARRIVAL                     TYPE.OF.ADMSN
                0.0000000                          0.0000000                          0.0000000
   TOTAL_COST_TO_HOSPITAL TOTAL.AMOUNT.BILLED.TO.THE.PATIENT                         CONCESSION
                0.0000000                          0.0000000                          0.0000000
   ACTUAL.RECEIVABLE.AMOUNT                TOTAL_LENGTH_OF_STAY                LENGTH_OF_STAY_ICU
                0.0000000                          0.0000000                          0.0000000
       LENGTH.OF.STAY..WARD                   IMPLANT.USED..Y.N.                            IMPLANT
                0.0000000                          0.0000000                          0.0000000
          COST_OF_IMPLANT
                0.0000000
```

Exhibit 2.2

AGE	BLOOD PRESSURE		HB		CREATININE	
	FEMALE	MALE	FEMALE	MALE	FEMALE	MALE
1 - 2	80/34 - 120/75	83/38 - 117/76	12.0 g/dl	12.0 g/dl	0.3 to 0.7 mg/dL	0.3 to 0.7 mg/dL
3	100/59	100/61	12.5 g/dl	12.5 g/dl	0.5 to 1.0 mg/dL	0.5 to 1.0 mg/dL
4	102/62	101/64	12.5 g/dl	12.5 g/dl	0.5 to 1.0 mg/dL	0.5 to 1.0 mg/dL
5	104/65	103/66	12.5 g/dl	12.5 g/dl	0.5 to 1.0 mg/dL	0.5 to 1.0 mg/dL
6	105/68	104/68	12.5 g/dl	12.5 g/dl	0.5 to 1.0 mg/dL	0.5 to 1.0 mg/dL
7	106/70	106/69	13.5 g/dl	13.5 g/dl	0.5 to 1.0 mg/dL	0.5 to 1.0 mg/dL
8	107/71	108/71	13.5 g/dl	13.5 g/dl	0.5 to 1.0 mg/dL	0.5 to 1.0 mg/dL
9	109/72	110/72	13.5 g/dl	13.5 g/dl	0.5 to 1.0 mg/dL	0.5 to 1.0 mg/dL
10	111/73	112/73	13.5 g/dl	13.5 g/dl	0.5 to 1.0 mg/dL	0.5 to 1.0 mg/dL
11	113/74	114/74	13.5 g/dl	13.5 g/dl	0.5 to 1.0 mg/dL	0.5 to 1.0 mg/dL
12	115/74	116/75	13.5 g/dl	13.5 g/dl	0.5 to 1.0 mg/dL	0.5 to 1.0 mg/dL
13	117/75	117/76	14.0 g/dl	14.5 g/dl	0.5 to 1.0 mg/dL	0.5 to 1.0 mg/dL
14	120/75	119/77	14.0 g/dl	14.5 g/dl	0.5 to 1.0 mg/dL	0.5 to 1.0 mg/dL
15	120/76	120/78	14.0 g/dl	14.5 g/dl	0.5 to 1.0 mg/dL	0.5 to 1.0 mg/dL
16	120/78	120/78	14.0 g/dl	14.5 g/dl	0.5 to 1.0 mg/dL	0.5 to 1.0 mg/dL
17	120/80	120/78	14.0 g/dl	14.5 g/dl	0.5 to 1.0 mg/dL	0.5 to 1.0 mg/dL
18	120/80	120/80	14.0 g/dl	14.5 g/dl	0.5 to 1.0 mg/dL	0.5 to 1.0 mg/dL
19-24	120/79	120/79	14.0 g/dl	15.5 g/dl	0.6 to 1.1 mg/dL	0.9 to 1.3 mg/dL
25-29	120/80	121/80	14.0 g/dl	15.5 g/dl	0.6 to 1.1 mg/dL	0.9 to 1.3 mg/dL
30-35	122/81	123/82	14.0 g/dl	15.5 g/dl	0.6 to 1.1 mg/dL	0.9 to 1.3 mg/dL
36-39	123/82	124/83	14.0 g/dl	15.5 g/dl	0.6 to 1.1 mg/dL	0.9 to 1.3 mg/dL
40-45	124/83	125/83	14.0 g/dl	15.5 g/dl	0.6 to 1.1 mg/dL	0.9 to 1.3 mg/dL
46-49	126/84	127/84	14.0 g/dl	15.5 g/dl	0.6 to 1.1 mg/dL	0.9 to 1.3 mg/dL
50-55	129/85	128/85	14.0 g/dl	15.5 g/dl	0.6 to 1.1 mg/dL	0.9 to 1.3 mg/dL
56-59	130/86	131/87	14.0 g/dl	15.5 g/dl	0.6 to 1.1 mg/dL	0.9 to 1.3 mg/dL
60+	134/84	135/88	14.0 g/dl	15.5 g/dl	0.6 to 1.1 mg/dL	0.9 to 1.3 mg/dL

GOA INSTITUTE OF MANAGEMENT
PGDM (Big Data Analytics)
Predictive and Prescriptive Analytics
2020-21

Exhibit 2.3

SL.	AGE	GENDER	MARITAL.STATUS	KEY.COMPLAINTS..CODE	BODY_WEIGHT	BODY_HEIGHT	HR.PULSE	BP..HIGH	BP.LOW
158	0.03	M	UNMARRIED	other- respiratory	2	45	120	NA	N
183	0.42	M	UNMARRIED	other- heart	5	66	100	NA	N
87	0.58	F	UNMARRIED	other- respiratory	6	57	150	NA	N
212	0.67	F	UNMARRIED	ACHD	2	47	134	NA	N
36	0.83	M	UNMARRIED	other- heart	6	68	120	NA	N
37	0.83	M	MARRIED	CAD-TVD	78	173	82	130	8
225	0.92	F	UNMARRIED	other-tertalogy	6	76	90	NA	N
230	0.92	M	UNMARRIED	PM-VSD	6	76	130	NA	N
50	1.00	M	UNMARRIED	other-nervous	5	66	100	100	7

Exhibit 2.4

SL.	AGE	GENDER	MARITAL.STATUS	KEY.COMPLAINTS..CODE	BODY_WEIGHT	BODY_HEIGHT	HR.PULSE	BP..HIGH	BP.LOW
158	0.03	M	UNMARRIED	other- respiratory	2	45	120	NA	N
183	0.42	M	UNMARRIED	other- heart	5	66	100	NA	N
87	0.58	F	UNMARRIED	other- respiratory	6	57	150	NA	N
212	0.67	F	UNMARRIED	ACHD	2	47	134	NA	N
36	0.83	M	UNMARRIED	other- heart	6	68	120	NA	N
37	0.83	M	MARRIED	CAD-TVD	78	173	82	130	8
225	0.92	F	UNMARRIED	other-tertalogy	6	76	90	NA	N
230	0.92	M	UNMARRIED	PM-VSD	6	76	130	NA	N
50	1.00	M	UNMARRIED	other-nervous	5	66	100	100	7

Exhibit 2.5

SL.	AGE	GENDER	MARITAL.STATUS	KEY.COMPLAINTS..CODE	BODY_WEIGHT	BODY_HEIGHT	HR.PULSE	BP..HIGH
14	64.00	M	MARRIED		56	168	105	130
40	67.00	M	MARRIED		57	167	90	120
44	50.00	M	MARRIED		65	155	59	120
46	78.00	M	MARRIED		48	158	88	120
47	39.00	F	MARRIED		77	153	86	130
48	64.00	M	MARRIED		68	162	60	130
49	53.00	M	MARRIED		55	156	80	140
51	55.00	M	MARRIED		78	163	100	140
52	56.00	M	MARRIED		56	162	82	150
54	48.00	M	MARRIED		64	158	74	120
55	53.00	M	MARRIED		59	159	68	130
56	69.00	M	MARRIED		56	166	84	120
58	10.00	M	UNMARRIED		6	64	96	87
59	12.00	F	UNMARRIED		32	149	82	100
60	10.00	F	UNMARRIED		23	137	90	90
61	14.00	F	UNMARRIED		49	149	111	100
62	7.00	M	UNMARRIED		19	107	100	103
63	13.00	M	UNMARRIED		22	133	90	110
65	11.00	M	UNMARRIED		26	140	90	NA
67	33.00	F	MARRIED		63	147	68	120
68	21.00	F	UNMARRIED		51	153	74	110

GOA INSTITUTE OF MANAGEMENT
PGDM (Big Data Analytics)
Predictive and Prescriptive Analytics
2020-21

Exhibit 2.7

KEY.COMPLAINTS..CODE	BODY_WEIGHT	BODY_HEIGHT	HR.PULSE	BP..HIGH	BP.LOW	RR	PAST.MEDICAL.HISTORY.CODE	HB
other- heart	8	80	112	89	56	30		25
other- heart	11	76	102	102	64	28		13
other- heart	41	152	88	110	70	20		13
other- heart	11	93	104	96	50	24		16
other- heart	60	185	90	120	90	22		10
other- heart	5	71	104	100	60	24		12
other- heart	46	62	96	140	90	20		5
other- heart	41	162	74	100	70	24		12
other- heart	18	117	83	98	70	24		9
other- heart	15	99	102	100	53	24		12
other- heart	18	118	106	130	80	24		13
other- heart	32	151	102	180	130	24		18
other- heart	16	120	98	90	50	24		20
other- heart	18	120	82	113	73	24		NA
other- heart	18	112	100	120	80	30		26
other- heart	7	68	112	80	50	24		12
other- heart	14	109	101	112	62	22		11
other- heart	9	78	100	84	69	30		13
other- heart	15	99	80	100	70	20		19
other- heart	15	105	110	NA	NA	32		12

Exhibit 2.6

BODY_WEIGHT	BODY_HEIGHT	HR.PULSE	BP..HIGH	BP.LOW	RR	PAST.MEDICAL.HISTORY.CODE	HB	UREA	CREATININE	MODE.OF.ARRIVAL	STATE.AT.THE.TIME.OF.ARRIVAL
All	All	All	All	All	All		All	All	All	All	All
56	168	105	130	80	22	Diabetes1	[illegible]	2	1.0	WALKED IN	ALERT
57	167	90	120	80	24	Diabetes2	[illegible]	24	1.0	WALKED IN	ALERT
65	155	59	120	70	[illegible]	hypertension1	16	26	1.0	AMBULANCE	ALERT
48	158	88	120	70	20	Hypertension1	12	18	1.0	WALKED IN	ALERT
77	153	86	130	80	26	hypertension2	13	21	1.0	WALKED IN	ALERT
68	162	60	130	90	24	hypertension3	14	22	1.0	WALKED IN	ALERT
55	156	80	140	80	20	other	13	20	1.0	WALKED IN	ALERT
78	163	100	140	90	22		15	24	1.0	WALKED IN	ALERT
56	162	82	150	80	24	hypertension2	10	42	3.0	TRANSFERRED	ALERT
64	158	74	120	70	22	hypertension2	12	28	1.0	WALKED IN	ALERT
59	159	68	130	80	16	hypertension3	14	19	1.0	WALKED IN	ALERT
56	166	84	120	70	24	Hypertension1	8	16	1.0	WALKED IN	ALERT
6	64	96	87	57	26	other	12	67	1.0	WALKED IN	ALERT
32	149	82	100	50	24		9	15	NA	WALKED IN	ALERT
23	137	90	90	60	22		11	18	NA	WALKED IN	ALERT

column 12: factor with 8 levels

Exhibit 2.8

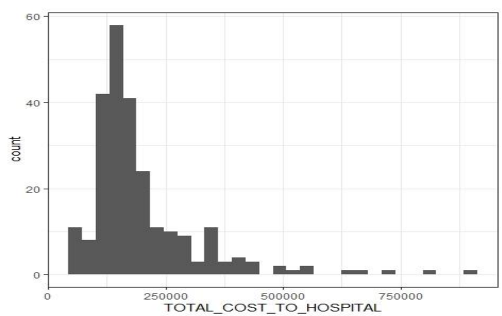
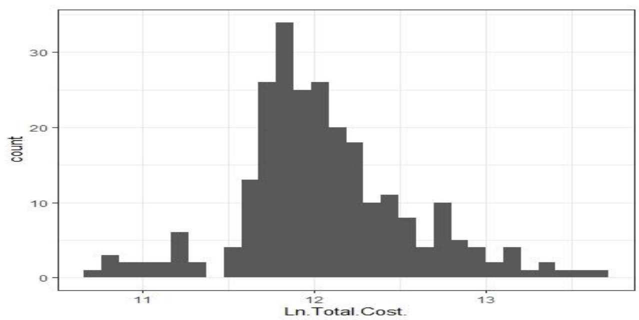


Exhibit 2.9



GOA INSTITUTE OF MANAGEMENT
PGDM (Big Data Analytics)
Predictive and Prescriptive Analytics
2020-21

Exhibit 2.10

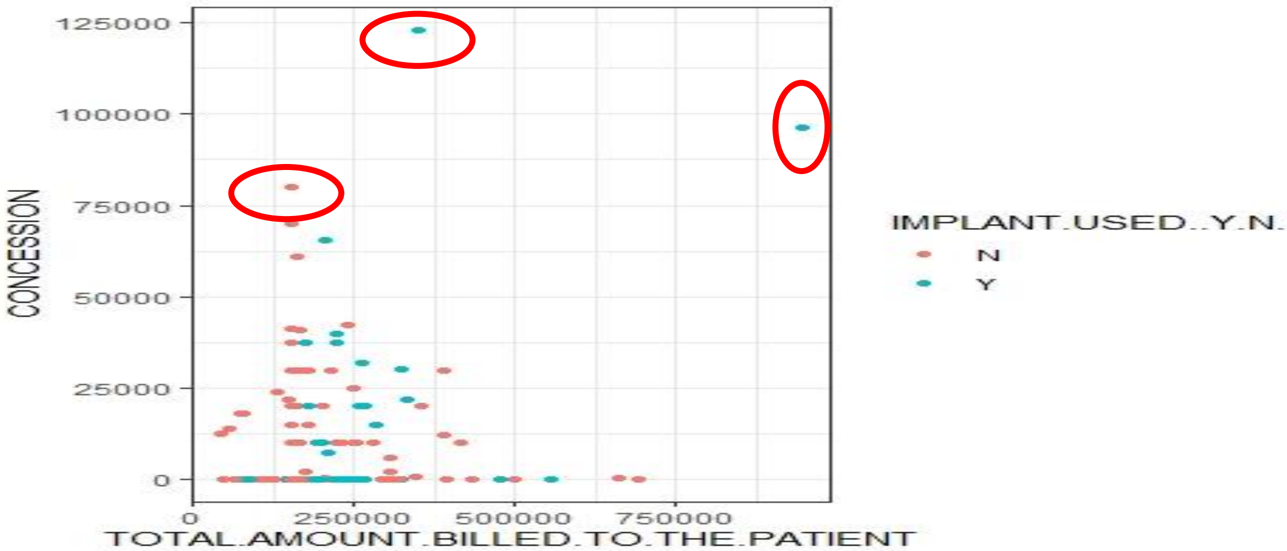


Exhibit 2.11

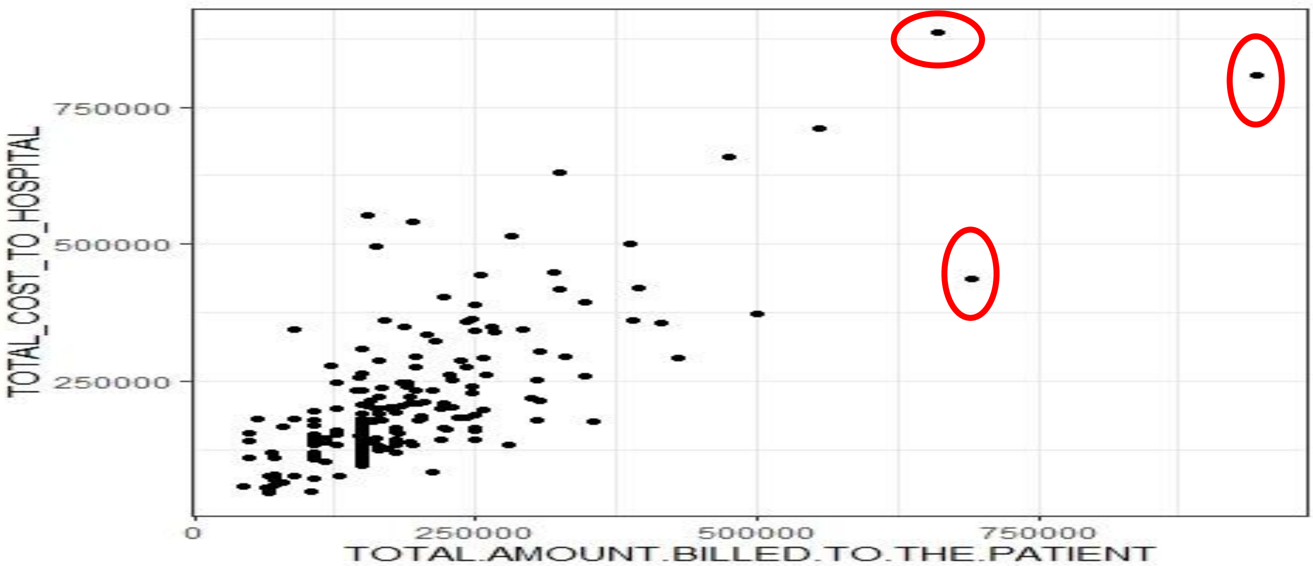
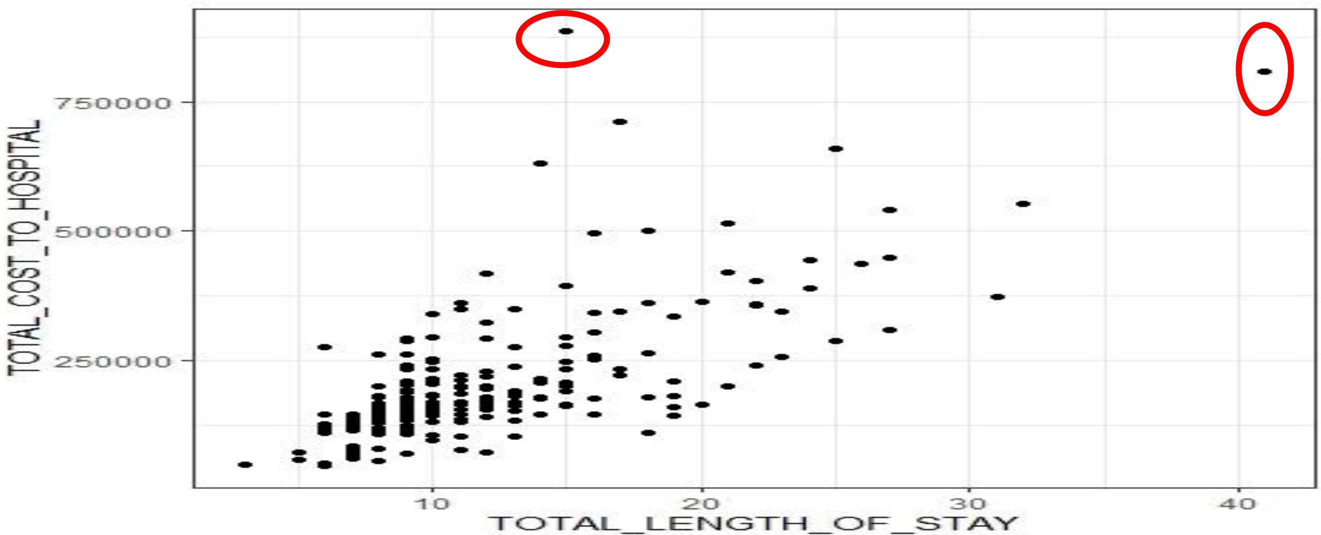


Exhibit 2.12



GOA INSTITUTE OF MANAGEMENT
PGDM (Big Data Analytics)
Predictive and Prescriptive Analytics
2020-21

Exhibit 2.13

	SL.	AGE	GENDER	MARITAL.STATUS	KEY.COMPLAINTS..CODE	BODY_WEIGHT	BODY_HEIGHT	HR.PULSE	BP..HIGH
158	158	0.03	M	UNMARRIED	other- respiratory	2	45	120	N
183	183	0.42	M	UNMARRIED	other- heart	5	66	100	N
87	87	0.58	F	UNMARRIED	other- respiratory	6	57	150	N
212	212	0.67	F	UNMARRIED	ACHD	2	47	134	N
36	36	0.83	M	UNMARRIED	other- heart	6	68	120	N
37	37	0.83	M	MARRIED	CAD-TVD	78	173	82	1
225	225	0.92	F	UNMARRIED	other-tertalogy	6	76	90	N
230	230	0.92	M	UNMARRIED	PM-VSD	6	76	130	N
50	50	1.00	M	UNMARRIED	other-nervous	5	66	100	1

GOA INSTITUTE OF MANAGEMENT
PGDM (Big Data Analytics)
Predictive and Prescriptive Analytics
2020-21

Exhibit 3.1

```
> modelQ3 <- lm(TOTAL_COST_TO_HOSPITAL~BODY_WEIGHT,mission_hospital)
> summary(modelQ3)

Call:
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

> 
```

Exhibit 4.1

	AGE	BODY_WEIGHT	BODY_HEIGHT	TOTAL_LENGTH_OF_STAY	LENGTH_OF_STAY_ICU	COST_OF_IMPLANT	TOTAL_COST_TO_HOSPITAL
AGE	1	0.85	0.72	0.22	0.35	0.18	0.42
BODY_WEIGHT	0.85	1	0.85	0.05	0.23	0.28	0.35
BODY_HEIGHT	0.72	0.85	1		0.14	0.28	0.29
TOTAL_LENGTH_OF_STAY	0.22	0.05		1	0.72	0.11	0.7
LENGTH_OF_STAY_ICU	0.35	0.23	0.14	0.72	1	0.16	0.82
COST_OF_IMPLANT	0.18	0.28	0.28	0.11	0.16	1	0.48
TOTAL_COST_TO_HOSPITAL	0.42	0.35	0.29	0.7	0.82	0.48	1

1 0.8 0.6 0.4 0.2 0 -0.2 -0.4 -0.6 -0.8 -1

GOA INSTITUTE OF MANAGEMENT
PGDM (Big Data Analytics)
Predictive and Prescriptive Analytics
2020-21

Exhibit 5.1

```
> mission_Q4 <- select(mission_hospital,SL.,AGE,BODY_WEIGHT,BODY_HEIGHT,TOTAL_LENGTH_OF_STAY,
+                      LENGTH_OF_STAY_ICU,COST_OF_IMPLANT,TOTAL_COST_TO_HOSPITAL)
> View(mission_Q4)
> mission_Q4$SL.<- NULL
> modelQ4 <- lm(TOTAL_COST_TO_HOSPITAL~.,mission_Q4)
> summary(modelQ4)

Call:
lm(formula = TOTAL_COST_TO_HOSPITAL ~ ., data = mission_Q4)

Residuals:
    Min      1Q  Median      3Q     Max
-195086  -21699    -945   20642  457682

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)         12469.3381 16028.3609   0.778   0.4374
AGE                   284.2982   245.3216   1.159   0.2477
BODY_WEIGHT           -96.2670   347.5705  -0.277   0.7820
BODY_HEIGHT           296.9352   153.6696   1.932   0.0545
TOTAL_LENGTH_OF_STAY 5546.1191   864.9432   6.412 7.51e-10 ***
LENGTH_OF_STAY_ICU  17903.3339  1234.1531  14.506  < 2e-16 ***
COST_OF_IMPLANT         1.9141     0.1549  12.356  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 49250 on 241 degrees of freedom
Multiple R-squared:  0.8425,	Adjusted R-squared:  0.8386
F-statistic: 214.8 on 6 and 241 DF,  p-value: < 2.2e-16

>
```

Exhibit 5.2

```
> modelQ4X <- lm(TOTAL_COST_TO_HOSPITAL~AGE+COST_OF_IMPLANT+LENGTH_OF_STAY_ICU+TOTAL_LENGTH_OF_STAY,mission_Q4)
> summary(modelQ4X)

Call:
lm(formula = TOTAL_COST_TO_HOSPITAL ~ AGE + COST_OF_IMPLANT +
    LENGTH_OF_STAY_ICU + TOTAL_LENGTH_OF_STAY, data = mission_Q4)

Residuals:
    Min      1Q  Median      3Q     Max
-194395  -21376    1047   20572  458951

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)          4.284e+04  8.632e+03   4.963 1.30e-06 ***
AGE                  5.397e+02  1.312e+02   4.115 5.31e-05 ***
COST_OF_IMPLANT      1.989e+00  1.498e-01  13.277  < 2e-16 ***
LENGTH_OF_STAY_ICU   1.778e+04  1.236e+03  14.383  < 2e-16 ***
TOTAL_LENGTH_OF_STAY 5.295e+03  8.466e+02   6.254 1.78e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 49560 on 243 degrees of freedom
Multiple R-squared:  0.8392,	Adjusted R-squared:  0.8365
F-statistic:   317 on 4 and 243 DF,  p-value: < 2.2e-16
```

GOA INSTITUTE OF MANAGEMENT
PGDM (Big Data Analytics)
Predictive and Prescriptive Analytics
2020-21

Exhibit 5.3

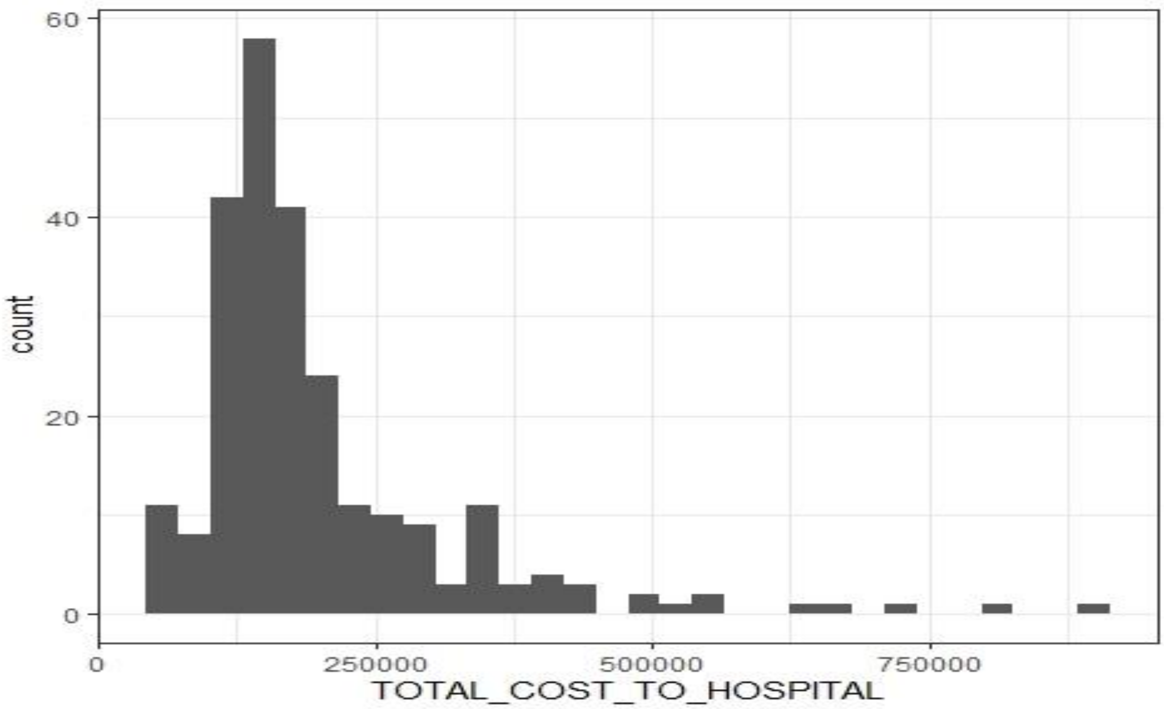


Exhibit 5.4

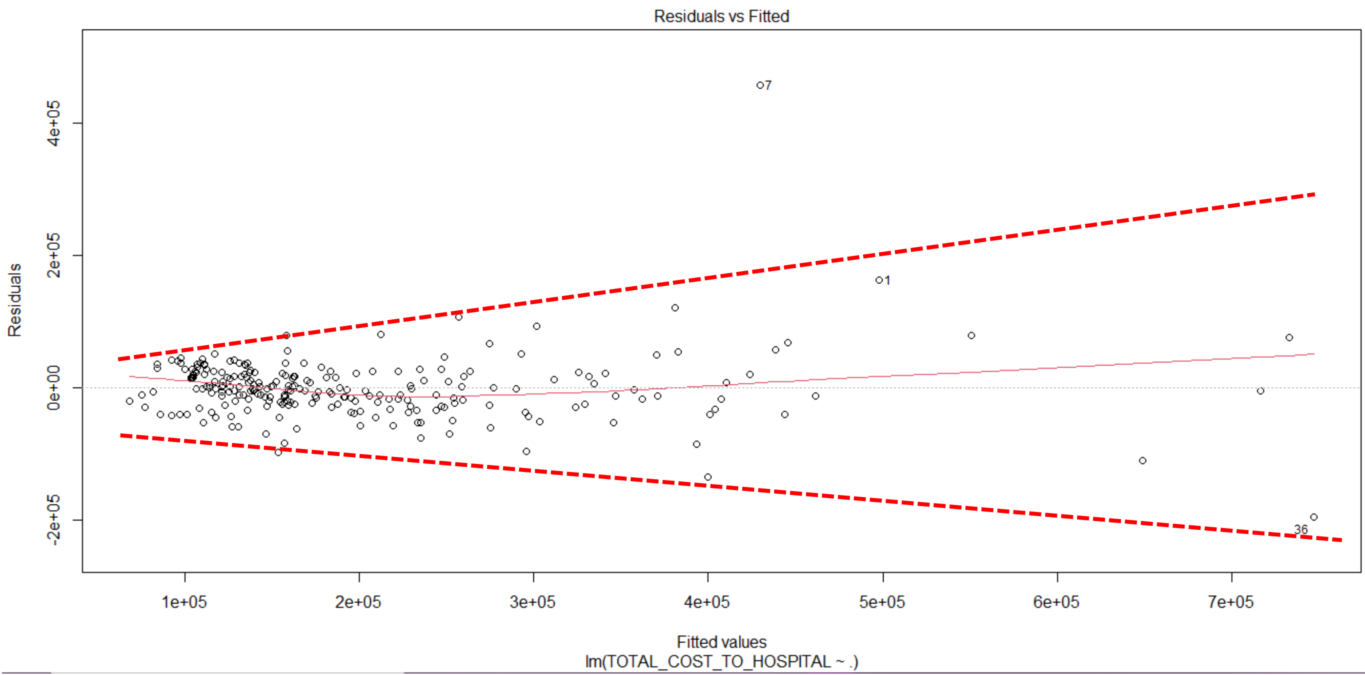


Exhibit 5.5

```
> ols_test_breusch_pagan(modelQ4X,rhs = TRUE,multiple = TRUE)

 Breusch Pagan Test for Heteroskedasticity
 -----------------------------------------
 Ho: the variance is constant
 Ha: the variance is not constant

                                Data
 -------------------------------------------------------------------------
 Response : TOTAL_COST_TO_HOSPITAL
 Variables: AGE COST_OF_IMPLANT LENGTH_OF_STAY_ICU TOTAL_LENGTH_OF_STAY

              Test Summary (Unadjusted p values)
 -----------------------------------------------------------
  Variable                   chi2        df          p
 -----------------------------------------------------------
  AGE                     64.3078481     1      1.064210e-15
  COST_OF_IMPLANT          0.1956134     1      6.582851e-01
  LENGTH_OF_STAY_ICU     470.4944279     1     2.501369e-104
  TOTAL_LENGTH_OF_STAY    86.9601624     1      1.107293e-20
 -----------------------------------------------------------
  simultaneous           571.1257518     4     2.746874e-122
 -----------------------------------------------------------
> |
```

Exhibit 5.6

```
> modelQ4U <- lm(Ln.Total.Cost.~.,mission_Q4U)
> summary(modelQ4U)

Call:
lm(formula = Ln.Total.Cost. ~ ., data = mission_Q4U)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.97165 -0.08191  0.03173  0.14851  0.88961 

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)    
(Intercept)          1.115e+01  8.132e-02 137.120  < 2e-16 ***
AGE                  4.876e-04  1.245e-03   0.392   0.6956    
BODY_WEIGHT          1.701e-03  1.763e-03   0.964   0.3358    
BODY_HEIGHT          1.454e-03  7.796e-04   1.865   0.0635 .  
TOTAL_LENGTH_OF_STAY 3.541e-02  4.388e-03   8.069 3.36e-14 ***
LENGTH_OF_STAY_ICU   4.933e-02  6.261e-03   7.879 1.14e-13 ***
COST_OF_IMPLANT      7.217e-06  7.860e-07   9.183  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2499 on 241 degrees of freedom
Multiple R-squared:  0.7615,	Adjusted R-squared:  0.7555 
F-statistic: 128.2 on 6 and 241 DF,  p-value: < 2.2e-16

> |
```

GOA INSTITUTE OF MANAGEMENT
PGDM (Big Data Analytics)
Predictive and Prescriptive Analytics
2020-21

Exhibit 5.7

```
> modelQ4UX <- lm(Ln.Total.Cost.~AGE+COST_OF_IMPLANT+LENGTH_OF_STAY_ICU+TOTAL_LENGTH_OF_STAY,mission_Q4U)
> summary(modelQ4UX)

Call:
lm(formula = Ln.Total.Cost. ~ AGE + COST_OF_IMPLANT + LENGTH_OF_STAY_ICU +
    TOTAL_LENGTH_OF_STAY, data = mission_Q4U)

Residuals:
     Min       1Q   Median       3Q      Max
-1.02565 -0.06231  0.04062  0.15111  0.87457

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)          1.135e+01  4.428e-02 256.239  < 2e-16 ***
AGE                  3.381e-03  6.729e-04   5.025 9.78e-07 ***
COST_OF_IMPLANT      7.902e-06  7.687e-07  10.280  < 2e-16 ***
LENGTH_OF_STAY_ICU   4.923e-02  6.343e-03   7.761 2.33e-13 ***
TOTAL_LENGTH_OF_STAY 3.259e-02  4.343e-03   7.504 1.17e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2543 on 243 degrees of freedom
Multiple R-squared:  0.751,	Adjusted R-squared:  0.7469
F-statistic: 183.2 on 4 and 243 DF,  p-value: < 2.2e-16
```

Exhibit 5.8

```
> ols_test_breusch_pagan(modelQ4UX,rhs = TRUE,multiple = TRUE)

 Breusch Pagan Test for Heteroskedasticity
 -----------------------------------------
 Ho: the variance is constant
 Ha: the variance is not constant

                              Data
 ----------------------------------------------------------------------
 Response : Ln.Total.Cost.
 Variables: AGE COST_OF_IMPLANT LENGTH_OF_STAY_ICU TOTAL_LENGTH_OF_STAY

              Test Summary (Unadjusted p values)
 ----------------------------------------------------------
  Variable                    chi2        df        p
 ----------------------------------------------------------
  AGE                      0.329154986     1    5.661571e-01
  COST_OF_IMPLANT          0.004250998     1    9.480150e-01
  LENGTH_OF_STAY_ICU       9.516229024     1    2.036627e-03
  TOTAL_LENGTH_OF_STAY     0.357819162     1    5.497199e-01
 ----------------------------------------------------------
  simultaneous            30.103197728     4    4.663285e-06
 ----------------------------------------------------------
```

GOA INSTITUTE OF MANAGEMENT
PGDM (Big Data Analytics)
Predictive and Prescriptive Analytics
2020-21

Exhibit 5.9

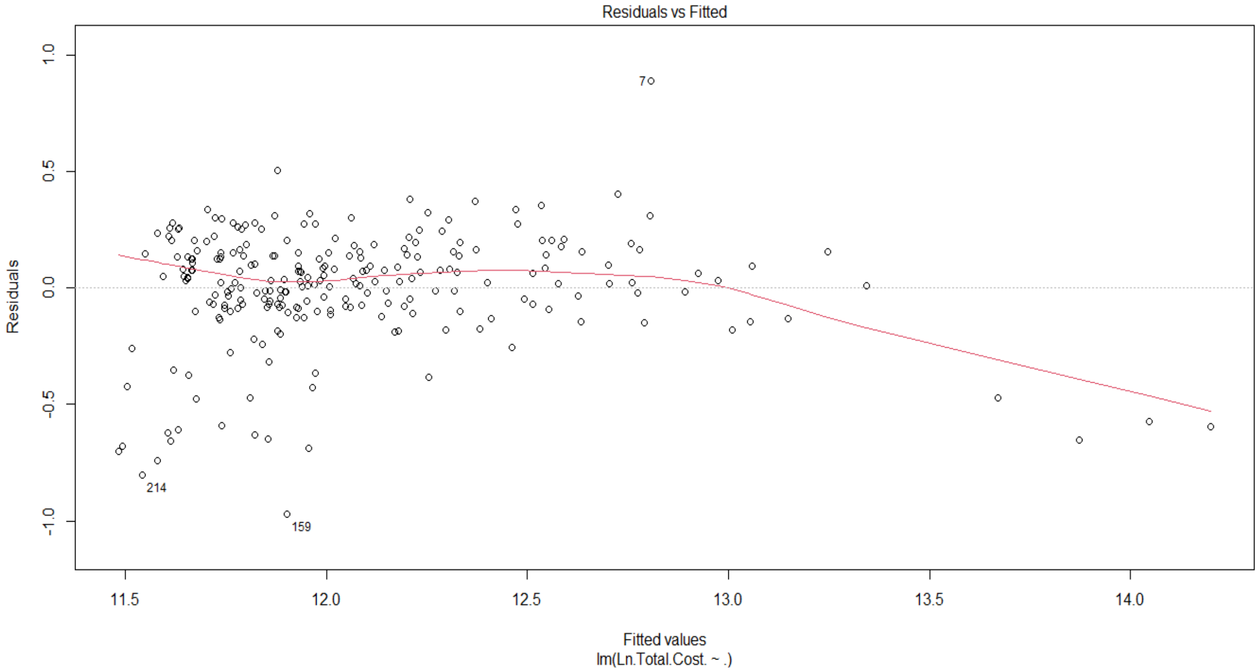


Exhibit 5.10

S.No.	Model	Independent Variables	Significant Variables	Adjusted R^2 Value
1	Model1	AGE	AGE ***	0.1488
2	Model2	AGE + BODY_WEIGHT	AGE **	0.1438
3	Model3	AGE + BODY_WEIGHT + BODY_HEIGHT	AGE **	0.1387
4	Model4	AGE + BODY_WEIGHT + BODY_HEIGHT + TOTAL_LENGTH_OF_STAY	BODY_WEIGHT * TOTAL_LENGTH_OF_STAY ***	0.5748
5	Model5	AGE + BODY_WEIGHT + BODY_HEIGHT + TOTAL_LENGTH_OF_STAY + LENGTH_OF_STAY_ICU	BODY_HEIGHT . TOTAL_LENGTH_OF_STAY *** LENGTH_OF_STAY_ICU ***	0.7367
6	Model6	AGE + BODY_WEIGHT + BODY_HEIGHT + TOTAL_LENGTH_OF_STAY + LENGTH_OF_STAY_ICU + COST_OF_IMPLANT	BODY_HEIGHT . TOTAL_LENGTH_OF_STAY *** LENGTH_OF_STAY_ICU *** COST_OF_IMPLANT ***	0.8386