# Regression Concepts Using R

Kumar Rahul

## In this exercise, we will use the patient data and understand the following:

1. Importing the datset from a csv file
2. Understanding the strucutre and summary of the data
3. Typecasting a variable to a proper data type
4. Creating derived variables and interaction variables
5. Analyzing the corelation amongst variables
6. Releveling the factor variable and understand its impact
7. Building the regression model using caret package
8. Writing the model equation and interpreting the model summary
9. Analyzing the statistics to ascertain the validity of the model

There are bugs/missing code in the entire exercise. The participants are expected to work upon them.

## Here are some useful links:

1. Refer link to know more about different ways of dummy variable coding
2. Read about interaction variable coding
3. Refer link to know about adding labels to factors
4. Refer link to relevel factor variables
5. Read about the issues in stepwise regression
6. The issues arising out of multi-colinearity is discussed here or here
7. The residual diagonstic can be interpreted from here
8. Read to understand the distinction between **outliers** and **influential cases**
9. Change NAs to a new label
10. Issues with rJava installation may get resolved by following link or by link

---

## Code starts here

We are going to use below mentioned libraries for demonstrating logistic regression:

```
library(stats)    #for regression
library(caret)    #for data partition
library(car)      #for VIF
library(sandwich) #for variance, covariance matrix
```

## Data Import and Manipulation

### 1. Importing a data set

*Give the correct path to the data*

```
raw.data <- read.csv("/Users/Rahul/Documents/Datasets/Mission Hospital-
Case Data.csv",
    header = TRUE, sep = ",", na.strings = c("", " ", "NA"))
```

Note that `echo = FALSE` parameter prevents printing the R code that generated the plot.

### 2a. Structure and Summary of the dataset

There are 175 NA values in Past Medical History Code. However, rather than treating these as missing values, it represents that there is no past medical history for these patients. These NA may be marked as "None". But while doing so, the code will give an error as we are trying to add a new level to factor variable (**raw.data$Past.MEDICAL.HISTORY.CODE**). In order to add a new level, first we will need to typecast this variable as a character variable, add a new level and then re-typecast them as Factor variable.

```
str(raw.data)

## 'data.frame':    250 obs. of  62 variables:
##  $ SL.                              : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ AGE                              : num  58 59 82 46 60 75 73 71
72 61 ...
##  $ GENDER                           : Factor w/ 2 levels "F","M": 2
2 2 2 2 2 2 2 2 ...
##  $ MALE                             : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Age.Gender                       : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ MARITAL.STATUS                   : Factor w/ 2 levels "MARRIED",
"UNMARRIED": 1 1 1 1 1 1 1 1 1 1 ...
##  $ UNMARRIED                        : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ KEY.COMPLAINTS..CODE             : Factor w/ 13 levels "ACHD","C
```

```
AD-DVD",..: 7 2 4 2 2 2 4 4 2 4 ...
##  $ ACHD                              : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ CAD.DVD                           : int  0 1 0 1 1 1 0 0 1 0 ...
##  $ CAD.SVD                           : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ CAD.TVD                           : int  0 0 1 0 0 0 1 1 0 1 ...
##  $ CAD.VSD                           : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ OS.ASD                            : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ other..heart                      : int  1 0 0 0 0 0 0 0 0 0 ...
##  $ other..respiratory                : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ other.general                     : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ other.nervous                     : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ other.tertalogy                   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ PM.VSD                            : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ RHD                               : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ BODY.WEIGHT                       : int  49 41 47 80 58 45 60 44
72 77 ...
##  $ Gender.Weight                     : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ BODY.HEIGHT                       : int  160 155 164 173 175 140
170 164 174 175 ...
##  $ Gender.Body.Height                : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ HR.PULSE                          : int  118 78 100 122 72 130 10
8 60 95 66 ...
##  $ BP..HIGH                          : int  100 70 110 110 180 215 1
60 130 100 140 ...
##  $ BP.LOW                            : int  80 50 80 80 100 140 90 9
0 50 90 ...
##  $ RR                                : int  32 28 20 24 18 42 24 22
25 22 ...
##  $ PAST.MEDICAL.HISTORY.CODE         : Factor w/ 7 levels "Diabetes1
","Diabetes2",..: NA NA 2 3 2 NA 2 NA 2 NA ...
##  $ Diabetes1                         : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Diabetes2                         : int  0 0 1 0 1 0 1 0 1 0 ...
##  $ hypertension1                     : int  0 0 0 1 0 0 0 0 0 0 ...
##  $ hypertension2                     : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ hypertension3                     : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ other                             : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ HB                                : int  11 11 12 12 10 12 15 10
10 14 ...
##  $ UREA                              : num  33 95 15 74 48 29 31 37
32 15 ...
##  $ CREATININE                        : num  0.8 1.7 0.8 1.5 1.9 1 1.
6 1.5 1.2 0.4 ...
##  $ MODE.OF.ARRIVAL                   : Factor w/ 3 levels "AMBULANCE
","TRANSFERRED",..: 1 1 3 1 1 1 3 3 1 3 ...
##  $ AMBULANCE                         : int  1 1 0 1 1 1 0 0 1 0 ...
##  $ TRANSFERRED                       : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ STATE.AT.THE.TIME.OF.ARRIVAL      : Factor w/ 2 levels "ALERT","C
ONFUSED": 1 1 1 1 1 1 1 1 1 1 ...
##  $ ALERT                             : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ TYPE.OF.ADMSN                     : Factor w/ 2 levels "ELECTIVE"
```

```
,"EMERGENCY": 2 2 1 2 2 2 1 2 2 1 ...
##  $ ELECTIVE                          : int  0 0 1 0 0 0 1 0 0 1 ...
##  $ TOTAL.COST.TO.HOSPITAL            : num  660293 809130 362231 629
990 444876 ...
##  $ Ln.Total.Cost.                    : num  13.4 13.6 12.8 13.4 13 .
..
##  $ TOTAL.AMOUNT.BILLED.TO.THE.PATIENT: int  474901 944819 390000 324
910 254673 499987 660504 248580 691297 247654 ...
##  $ CONCESSION                        : int  0 96422 30000 0 10000 0
504 0 0 0 ...
##  $ ACTUAL.RECEIVABLE.AMOUNT          : int  474901 848397 360000 324
910 244673 499987 660000 248580 691297 247654 ...
##  $ TOTAL.LENGTH.OF.STAY              : int  25 41 18 14 24 31 15 24
26 20 ...
##  $ LENGTH.OF.STAY...ICU              : int  12 20 9 13 12 9 15 11 9
4 ...
##  $ LENGTH.OF.STAY..WARD              : int  13 21 9 1 12 22 0 13 17
16 ...
##  $ IMPLANT.USED..Y.N.                : Factor w/ 2 levels "N","Y": 2
2 1 2 1 1 1 1 1 1 ...
##  $ IMPLANT                           : int  1 1 0 1 0 0 0 0 0 0 ...
##  $ COST.OF.IMPLANT                   : int  38000 39690 0 89450 0 0
0 0 0 0 ...
##  $ Y.hat                             : num  260518 262706 313011 234
272 264893 ...
##  $ APE                               : num  0.605 0.675 0.136 0.628
0.405 ...
##  $ X                                 : logi  NA NA NA NA NA NA ...
##  $ X.1                               : logi  NA NA NA NA NA NA ...
##  $ S.D                               : num  1.01e+05 NA 1.28 3.90e+0
5 NA ...
```

```r
summary(raw.data)
```

```
##       SL.              AGE         GENDER         MALE
##  Min.   :  1.00   Min.   : 0.03   F   : 82   Min.   :0.0000
##  1st Qu.: 62.75   1st Qu.: 6.00   M   :166   1st Qu.:0.0000
##  Median :124.50   Median :15.50   NA's:  2   Median :0.0000
##  Mean   :124.50   Mean   :28.88              Mean   :0.3306
##  3rd Qu.:186.25   3rd Qu.:55.00              3rd Qu.:1.0000
##  Max.   :248.00   Max.   :88.00              Max.   :1.0000
##  NA's   :2        NA's   :2                  NA's   :2
##    Age.Gender       MARITAL.STATUS   UNMARRIED
##  Min.   : 0.000   MARRIED  :108    Min.   :-0.8985
##  1st Qu.: 0.000   UNMARRIED:140    1st Qu.: 0.0000
##  Median : 0.000   NA's     :  2    Median : 1.0000
##  Mean   : 7.206                    Mean   : 0.5586
##  3rd Qu.: 4.250                    3rd Qu.: 1.0000
##  Max.   :78.000                    Max.   : 1.0000
##  NA's   :2                         NA's   :1
##    KEY.COMPLAINTS..CODE      ACHD            CAD.DVD
##  other- heart:55        Min.   :0.00000   Min.   :0.0000
```

```
##  CAD-DVD     :27        1st Qu.:0.00000   1st Qu.:0.0000
##  RHD         :26        Median :0.00000   Median :0.0000
##  CAD-TVD     :24        Mean   :0.07661   Mean   :0.1089
##  ACHD        :19        3rd Qu.:0.00000   3rd Qu.:0.0000
##  (Other)     :61        Max.   :1.00000   Max.   :1.0000
##  NA's        :38        NA's   :2         NA's   :2
##     CAD.SVD            CAD.TVD           CAD.VSD             OS.ASD
##  Min.   :0.000000   Min.   :0.00000   Min.   :0.000000   Min.   :0.00000
##  1st Qu.:0.000000   1st Qu.:0.00000   1st Qu.:0.000000   1st Qu.:0.00000
##  Median :0.000000   Median :0.00000   Median :0.000000   Median :0.00000
##  Mean   :0.008065   Mean   :0.09677   Mean   :0.004032   Mean   :0.06048
##  3rd Qu.:0.000000   3rd Qu.:0.00000   3rd Qu.:0.000000   3rd Qu.:0.00000
##  Max.   :1.000000   Max.   :1.00000   Max.   :1.000000   Max.   :1.00000
##  NA's   :2          NA's   :2         NA's   :2          NA's   :2
##   other..heart    other..respiratory other.general      other.nervous
##  Min.   :0.0000   Min.   :0.00000    Min.   :0.000000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.00000    1st Qu.:0.000000   1st Qu.:0.0000
##  Median :0.0000   Median :0.00000    Median :0.000000   Median :0.0000
##  Mean   :0.2218   Mean   :0.06048    Mean   :0.004032   Mean   :0.0121
##  3rd Qu.:0.0000   3rd Qu.:0.00000    3rd Qu.:0.000000   3rd Qu.:0.0000
##  Max.   :1.0000   Max.   :1.00000    Max.   :1.000000   Max.   :1.0000
##  NA's   :2        NA's   :2          NA's   :2          NA's   :2
##  other.tertalogy       PM.VSD             RHD           BODY.WEIGHT
##  Min.   :0.00000   Min.   :0.00000   Min.   :0.0000   Min.   : 2.00
##  1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.0000   1st Qu.:15.00
##  Median :0.00000   Median :0.00000   Median :0.0000   Median :41.00
##  Mean   :0.07258   Mean   :0.02419   Mean   :0.1048   Mean   :37.54
##  3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.0000   3rd Qu.:58.25
##  Max.   :1.00000   Max.   :1.00000   Max.   :1.0000   Max.   :85.00
##  NA's   :2         NA's   :2         NA's   :2        NA's   :2
##  Gender.Weight    BODY.HEIGHT    Gender.Body.Height    HR.PULSE
##  Min.   : 0.00   Min.   : 19.0   Min.   :  0.00     Min.   : 41.00
##  1st Qu.: 0.00   1st Qu.:105.0   1st Qu.:  0.00     1st Qu.: 78.00
##  Median : 0.00   Median :147.5   Median :  0.00     Median : 90.00
##  Mean   :10.51   Mean   :130.2   Mean   : 40.47     Mean   : 92.23
##  3rd Qu.:12.25   3rd Qu.:160.0   3rd Qu.: 81.00     3rd Qu.:104.00
##  Max.   :77.00   Max.   :185.0   Max.   :167.00     Max.   :155.00
##  NA's   :2       NA's   :2       NA's   :2          NA's   :2
##     BP..HIGH       BP.LOW             RR        PAST.MEDICAL.HISTORY.CODE
##  Min.   : 70   Min.   : 39.00   Min.   :12.00   hypertension1: 20
##  1st Qu.:100   1st Qu.: 60.00   1st Qu.:22.00   other        : 15
##  Median :110   Median : 70.00   Median :24.00   hypertension2: 13
##  Mean   :115   Mean   : 71.88   Mean   :23.54   Diabetes1    : 10
##  3rd Qu.:130   3rd Qu.: 80.00   3rd Qu.:24.00   Diabetes2    :  9
##  Max.   :215   Max.   :140.00   Max.   :42.00   (Other)      :  8
##  NA's   :25    NA's   :25       NA's   :2       NA's         :175
##    Diabetes1         Diabetes2       hypertension1     hypertension2
##  Min.   :0.00000   Min.   :0.00000   Min.   :0.00000   Min.   :0.00000
##  1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.00000
##  Median :0.00000   Median :0.00000   Median :0.00000   Median :0.00000
##  Mean   :0.04032   Mean   :0.03629   Mean   :0.09274   Mean   :0.05242
##  3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.00000
##  Max.   :1.00000   Max.   :1.00000   Max.   :1.00000   Max.   :1.00000
##  NA's   :2         NA's   :2         NA's   :2         NA's   :2
##  hypertension3         other               HB             UREA
```

```
##  Min.   :0.00000   Min.   :0.00000   Min.   : 5.00   Min.   :  2.00
##  1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:11.00   1st Qu.: 18.00
##  Median :0.00000   Median :0.00000   Median :12.00   Median : 22.00
##  Mean   :0.02016   Mean   :0.06048   Mean   :12.93   Mean   : 26.58
##  3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:14.00   3rd Qu.: 30.00
##  Max.   :1.00000   Max.   :1.00000   Max.   :26.00   Max.   :143.00
##  NA's   :2         NA's   :2         NA's   :4       NA's   :15
##    CREATININE       MODE.OF.ARRIVAL   AMBULANCE      TRANSFERRED
##  Min.   :0.100   AMBULANCE  : 30    Min.   :0.000   Min.   :0.00000
##  1st Qu.:0.300   TRANSFERRED:  4    1st Qu.:0.000   1st Qu.:0.00000
##  Median :0.700   WALKED IN  :214    Median :0.000   Median :0.00000
##  Mean   :0.747   NA's       :  2    Mean   :0.121   Mean   :0.01613
##  3rd Qu.:1.000                      3rd Qu.:0.000   3rd Qu.:0.00000
##  Max.   :5.200                      Max.   :1.000   Max.   :1.00000
##  NA's   :35                         NA's   :2       NA's   :2
##  STATE.AT.THE.TIME.OF.ARRIVAL     ALERT          TYPE.OF.ADMSN
##  ALERT   :247                 Min.   :0.000   ELECTIVE :216
##  CONFUSED:  1                 1st Qu.:1.000   EMERGENCY: 32
##  NA's    :  2                 Median :1.000   NA's     :  2
##                               Mean   :0.996
##                               3rd Qu.:1.000
##                               Max.   :1.000
##                               NA's   :2
##     ELECTIVE     TOTAL.COST.TO.HOSPITAL Ln.Total.Cost.
##  Min.   :0.000   Min.   : 46093         Min.   :10.74
##  1st Qu.:1.000   1st Qu.:131653         1st Qu.:11.79
##  Median :1.000   Median :162660         Median :12.00
##  Mean   :0.871   Mean   :198723         Mean   :12.06
##  3rd Qu.:1.000   3rd Qu.:220614         3rd Qu.:12.30
##  Max.   :1.000   Max.   :887350         Max.   :13.70
##  NA's   :2       NA's   :2              NA's   :2
##  TOTAL.AMOUNT.BILLED.TO.THE.PATIENT   CONCESSION
##  Min.   : 43641                     Min.   :     0
##  1st Qu.:150000                     1st Qu.:     0
##  Median :150000                     Median : 10000
##  Mean   :182721                     Mean   : 17643
##  3rd Qu.:202638                     3rd Qu.: 37500
##  Max.   :944819                     Max.   :123132
##  NA's   :2                          NA's   :2
##  ACTUAL.RECEIVABLE.AMOUNT TOTAL.LENGTH.OF.STAY LENGTH.OF.STAY...ICU
##  Min.   : 31000           Min.   : 3.00        Min.   : 0.000
##  1st Qu.:112500           1st Qu.: 8.00        1st Qu.: 1.000
##  Median :122400           Median :10.00        Median : 2.000
##  Mean   :167894           Mean   :11.61        Mean   : 3.476
##  3rd Qu.:197000           3rd Qu.:13.00        3rd Qu.: 4.000
##  Max.   :848397           Max.   :41.00        Max.   :30.000
##  NA's   :2                NA's   :2            NA's   :2
##  LENGTH.OF.STAY..WARD IMPLANT.USED..Y.N.    IMPLANT       COST.OF.IMPLANT
##  Min.   : 0.000       N   :199           Min.   :0.0000   Min.   :     0
##  1st Qu.: 6.000       Y   : 49           1st Qu.:0.0000   1st Qu.:     0
##  Median : 7.000       NA's:  2           Median :0.0000   Median :     0
##  Mean   : 8.153                          Mean   :0.1976   Mean   :  8544
##  3rd Qu.:10.000                          3rd Qu.:0.0000   3rd Qu.:     0
##  Max.   :22.000                          Max.   :1.0000   Max.   :196848
##  NA's   :2                               NA's   :2        NA's   :2
```

```
##      Y.hat             APE              X             X.1
##  Min.   :133733   Min.   :0.000013   Mode:logical   Mode:logical
##  1st Qu.:146784   1st Qu.:0.125760   NA's:250       NA's:250
##  Median :167562   Median :0.280740
##  Mean   :196827   Mean   :0.417690
##  3rd Qu.:253957   3rd Qu.:0.550740
##  Max.   :326134   Max.   :4.287823
##  NA's   :2        NA's   :1
##       S.D
##  Min.   :1.3
##  1st Qu.:50550.6
##  Median :101100.0
##  Mean   :163728.2
##  3rd Qu.:245591.7
##  Max.   :390083.4
##  NA's   :247
```

```
raw.data$PAST.MEDICAL.HISTORY.CODE[raw.data$PAST.MEDICAL.HISTORY.CODE =
= "Hypertension1"] <- "hypertension1"

raw.data$PAST.MEDICAL.HISTORY.CODE <- as.character(raw.data$PAST.MEDICA
L.HISTORY.CODE)

raw.data$PAST.MEDICAL.HISTORY.CODE[is.na(raw.data$PAST.MEDICAL.HISTORY.
CODE)] <- "None"

raw.data$PAST.MEDICAL.HISTORY.CODE <- as.factor(raw.data$PAST.MEDICAL.H
ISTORY.CODE)
```

Create a new data frame and store the raw data copy. This is being done to have a copy of the raw data intact for further manipulation if needed.

```
new.data <- raw.data[, c(-1, -4, -5, -7, -9:-21, -23, -25, -31:-36, -41
, -42,
    -44, -46, -48, -56, -58:-62)]
new.data <- na.omit(new.data)  # listwise deletion of missing
```

## 3a. Correlation among Variables

From the numeric attribute in the data, it will of interest to analyze the variables which are corelated to each other. High corelation amongst variable may result in the issue of **multi-colinearity** in the model.

```
correlationMatrix <- cor(new.data[, c(1, 7:10, 12:14, 18:24, 26)])
print(correlationMatrix)
```

```
##                                          AGE     HR.PULSE    BP..HIGH
## AGE                                1.00000000 -0.451244005  0.58656780
## HR.PULSE                          -0.45124400  1.000000000 -0.29163412
## BP..HIGH                           0.58656780 -0.291634124  1.00000000
## BP.LOW                             0.46545550 -0.207449219  0.77298853
## RR                                -0.23480792  0.373233721 -0.08309698
## HB                                -0.21849870  0.099654811 -0.08392965
```

```
## UREA                                 0.28568989 -0.024115762  0.09639492
## CREATININE                           0.70849144 -0.334538256  0.44300126
## TOTAL.COST.TO.HOSPITAL               0.49918592 -0.060194555  0.21756095
## TOTAL.AMOUNT.BILLED.TO.THE.PATIENT   0.49932971 -0.057115599  0.22629958
## CONCESSION                          -0.38706554  0.199744235 -0.29482834
## ACTUAL.RECEIVABLE.AMOUNT             0.54955029 -0.103888398  0.28100749
## TOTAL.LENGTH.OF.STAY                 0.34517109  0.009432666  0.12161925
## LENGTH.OF.STAY...ICU                 0.49472755 -0.080920600  0.18986251
## LENGTH.OF.STAY..WARD                -0.01321377  0.097867560 -0.02581442
## COST.OF.IMPLANT                      0.14886888 -0.044193648 -0.01621976
##                                          BP.LOW          RR          HB
## AGE                                  0.465455500 -0.23480792 -0.21849870
## HR.PULSE                            -0.207449219  0.37323372  0.09965481
## BP..HIGH                             0.772988535 -0.08309698 -0.08392965
## BP.LOW                               1.000000000 -0.01569492  0.03468884
## RR                                  -0.015694922  1.00000000  0.03551983
## HB                                   0.034688841  0.03551983  1.00000000
## UREA                                 0.043500316  0.06318983 -0.09670059
## CREATININE                           0.319224146 -0.15830983 -0.22771802
## TOTAL.COST.TO.HOSPITAL               0.211650056  0.04572571 -0.09422928
## TOTAL.AMOUNT.BILLED.TO.THE.PATIENT   0.199455448  0.06994042 -0.10141016
## CONCESSION                          -0.265444201  0.19567060  0.17308650
## ACTUAL.RECEIVABLE.AMOUNT             0.262555455  0.03910597 -0.11850792
## TOTAL.LENGTH.OF.STAY                 0.107979390  0.17024882 -0.02483995
## LENGTH.OF.STAY...ICU                 0.141540924  0.05138801 -0.13113079
## LENGTH.OF.STAY..WARD                 0.007833746  0.19557658  0.10441442
## COST.OF.IMPLANT                      0.061072583  0.05194928 -0.07064192
##                                            UREA  CREATININE
## AGE                                  0.28568989  0.70849144
## HR.PULSE                            -0.02411576 -0.33453826
## BP..HIGH                             0.09639492  0.44300126
## BP.LOW                               0.04350032  0.31922415
## RR                                   0.06318983 -0.15830983
## HB                                  -0.09670059 -0.22771802
## UREA                                 1.00000000  0.63917958
## CREATININE                           0.63917958  1.00000000
## TOTAL.COST.TO.HOSPITAL               0.28068028  0.51605814
## TOTAL.AMOUNT.BILLED.TO.THE.PATIENT   0.28324263  0.49946442
## CONCESSION                          -0.07309794 -0.27399988
## ACTUAL.RECEIVABLE.AMOUNT             0.28301870  0.52374603
## TOTAL.LENGTH.OF.STAY                 0.23601057  0.35459975
## LENGTH.OF.STAY...ICU                 0.25439972  0.48685662
## LENGTH.OF.STAY..WARD                 0.08392070  0.01665721
## COST.OF.IMPLANT                      0.24741685  0.19856159
##                                     TOTAL.COST.TO.HOSPITAL
## AGE                                              0.49918592
## HR.PULSE                                        -0.06019455
## BP..HIGH                                         0.21756095
## BP.LOW                                           0.21165006
## RR                                               0.04572571
## HB                                              -0.09422928
## UREA                                             0.28068028
## CREATININE                                       0.51605814
## TOTAL.COST.TO.HOSPITAL                           1.00000000
## TOTAL.AMOUNT.BILLED.TO.THE.PATIENT               0.79971528
```

```
## CONCESSION                                    -0.08280661
## ACTUAL.RECEIVABLE.AMOUNT                       0.77012057
## TOTAL.LENGTH.OF.STAY                           0.69772333
## LENGTH.OF.STAY...ICU                           0.84745307
## LENGTH.OF.STAY..WARD                           0.14441239
## COST.OF.IMPLANT                                0.47986318
##                                    TOTAL.AMOUNT.BILLED.TO.THE.PATIENT
## AGE                                                        0.49932971
## HR.PULSE                                                  -0.05711560
## BP..HIGH                                                   0.22629958
## BP.LOW                                                     0.19945545
## RR                                                         0.06994042
## HB                                                        -0.10141016
## UREA                                                       0.28324263
## CREATININE                                                 0.49946442
## TOTAL.COST.TO.HOSPITAL                                     0.79971528
## TOTAL.AMOUNT.BILLED.TO.THE.PATIENT                         1.00000000
## CONCESSION                                                 0.07128904
## ACTUAL.RECEIVABLE.AMOUNT                                   0.93057489
## TOTAL.LENGTH.OF.STAY                                       0.63274839
## LENGTH.OF.STAY...ICU                                       0.64058348
## LENGTH.OF.STAY..WARD                                       0.25678908
## COST.OF.IMPLANT                                            0.33145494
##                                     CONCESSION ACTUAL.RECEIVABLE.AMOUNT
## AGE                                -0.38706554                0.54955029
## HR.PULSE                            0.19974424               -0.10388840
## BP..HIGH                           -0.29482834                0.28100749
## BP.LOW                             -0.26544420                0.26255546
## RR                                  0.19567060                0.03910597
## HB                                  0.17308650               -0.11850792
## UREA                               -0.07309794                0.28301870
## CREATININE                         -0.27399988                0.52374603
## TOTAL.COST.TO.HOSPITAL             -0.08280661                0.77012057
## TOTAL.AMOUNT.BILLED.TO.THE.PATIENT  0.07128904                0.93057489
## CONCESSION                          1.00000000               -0.11758682
## ACTUAL.RECEIVABLE.AMOUNT           -0.11758682                1.00000000
## TOTAL.LENGTH.OF.STAY                0.01068904                0.61237607
## LENGTH.OF.STAY...ICU               -0.08786860                0.64942890
## LENGTH.OF.STAY..WARD                0.10330812                0.21882633
## COST.OF.IMPLANT                    -0.11763011                0.32354920
##                                    TOTAL.LENGTH.OF.STAY
## AGE                                         0.345171087
## HR.PULSE                                    0.009432666
## BP..HIGH                                    0.121619250
## BP.LOW                                      0.107979390
## RR                                          0.170248825
## HB                                         -0.024839945
## UREA                                        0.236010569
## CREATININE                                  0.354599755
## TOTAL.COST.TO.HOSPITAL                      0.697723335
## TOTAL.AMOUNT.BILLED.TO.THE.PATIENT          0.632748391
## CONCESSION                                  0.010689039
## ACTUAL.RECEIVABLE.AMOUNT                    0.612376067
## TOTAL.LENGTH.OF.STAY                        1.000000000
## LENGTH.OF.STAY...ICU                        0.721035337
```

```
## LENGTH.OF.STAY..WARD                          0.707134187
## COST.OF.IMPLANT                               0.112062033
##                                     LENGTH.OF.STAY...ICU
## AGE                                           0.49472755
## HR.PULSE                                     -0.08092060
## BP..HIGH                                      0.18986251
## BP.LOW                                        0.14154092
## RR                                            0.05138801
## HB                                           -0.13113079
## UREA                                          0.25439972
## CREATININE                                    0.48685662
## TOTAL.COST.TO.HOSPITAL                        0.84745307
## TOTAL.AMOUNT.BILLED.TO.THE.PATIENT            0.64058348
## CONCESSION                                   -0.08786860
## ACTUAL.RECEIVABLE.AMOUNT                      0.64942890
## TOTAL.LENGTH.OF.STAY                          0.72103534
## LENGTH.OF.STAY...ICU                          1.00000000
## LENGTH.OF.STAY..WARD                          0.02179490
## COST.OF.IMPLANT                               0.18278343
##                                     LENGTH.OF.STAY..WARD COST.OF.IMPLANT
## AGE                                         -0.013213772     0.14886888
## HR.PULSE                                     0.097867560    -0.04419365
## BP..HIGH                                    -0.025814415    -0.01621976
## BP.LOW                                       0.007833746     0.06107258
## RR                                           0.195576581     0.05194928
## HB                                           0.104414424    -0.07064192
## UREA                                         0.083920703     0.24741685
## CREATININE                                   0.016657206     0.19856159
## TOTAL.COST.TO.HOSPITAL                       0.144412386     0.47986318
## TOTAL.AMOUNT.BILLED.TO.THE.PATIENT           0.256789081     0.33145494
## CONCESSION                                   0.103308121    -0.11763011
## ACTUAL.RECEIVABLE.AMOUNT                     0.218826327     0.32354920
## TOTAL.LENGTH.OF.STAY                         0.707134187     0.11206203
## LENGTH.OF.STAY...ICU                         0.021794904     0.18278343
## LENGTH.OF.STAY..WARD                         1.000000000    -0.02250497
## COST.OF.IMPLANT                             -0.022504973     1.00000000
```

```r
# find attributes that are highly corrected (ideally >0.7)
highlyCorrelated <- findCorrelation(correlationMatrix, cutoff = 0.7, na
mes = TRUE)
print(highlyCorrelated)
```

```
## [1] "AGE"                          "ACTUAL.RECEIVABLE.AMOUNT"
## [3] "TOTAL.COST.TO.HOSPITAL"       "LENGTH.OF.STAY...ICU"
## [5] "TOTAL.LENGTH.OF.STAY"         "BP..HIGH"
```

## 3b. Derived variables

Deriving BMI to drop of Weight and Height as variables. Both of them where highly corelated to age. Droping Cretanine as a variable as it is highly corleated to age.

```
<Missing code to add a new variable named BMI.

BMI is weight (Kg)/(height in meter)^2>
```

```
<Create an interaction variable using "Implant used" and "Cost of Implant">
new.data$I_COST.OF.IMPLANT <- missing code
filter.data <- new.data[,c(-5:-6)]
```

### 3c. Relevel

By default, the base category/reference category selected is ordered alphabetically. In this code chunk we are just changing the base category for PAST.MEDICAL.HISTORY.CODE variable.

The base category can be releveled using the function **relevel()**.

```
filter.data$PAST.MEDICAL.HISTORY.CODE <- missing code
```

## 4. Create train and test datase vct

### Reserve 80% for *training* and 20% of *test*

*Correct the error in the below code chunk*

```
set.seed(2341)
trainIndex <- createDataPartition(filter.data$TOTAL.COST.TO.HOSPITAL, p
= 0.8,
    list = FALSE)
data.train <- filter.data[trainIndex, ]
data.test <- filter.data[trainIndex, ]
```

Transformation of variables may be needed to validate the model assumptions.

```
data.train$Log.Cost.Treatment <- log(data.train$TOTAL.COST.TO.HOSPITAL)
data.test$Log.Cost.Treatment <- log(data.test$TOTAL.COST.TO.HOSPITAL)
```

We can pull the specific attribute needed to build the model in another data frame. This agian is more of a hygine practice to not touch the **train** and **test** data set directly.

*Correct the error in the below code chunk*

```
reg.train.data <- as.data.frame(data.train[,c("AGE",
                                              "HR.PULSE",
                                              "BP..HIGH",
                                              "RR",
                                              "HB",
                                              "UREA",
                                              #"TOTAL.LENGTH.OF.STAY",
                                              "BMI",
                                              #"COST.OF.IMPLANT",
                                              #"IMPLANT.USED..Y.N.",
                                              "I_COST.OF.IMPLANT",
                                              "GENDER",
                                              "MARITL.STATUS",
```

```
                                                "KEY.COMPLAINTS..CODE",
                                                "PAST.MEDICAL.HISTORY.CODE
",
                                                "MODE.OF.ARRIVAL",
                                                "STATE.AT.THE.TIME.OF.ARRI
VAL",
                                                "TYPE.OF.ADMSN",
                                                "TOTAL.COST.TO.HOSPITAL"
                                                #"Log.Cost.Treatment"
)])
```

*Correct the error in the below code chunk*

```
reg.test.data <- as.data.frame(data.test[,c("AGE",
                                                "HR.PULSE",
                                                "BP..HIGH",
                                                "RR",
                                                "HB",
                                                "UREA",
                                                #"TOTAL.LENGTH.OF.STAY",
                                                "BMI",
                                                #"COST.OF.IMPLANT",
                                                #"IMPLANT.USED..Y.N.",
                                                "I_COST.OF.IMPLANT",
                                                "GENDER",
                                                "MARITAL.STATUS",
                                                "KEY.COMPLAINTS..CODE",
                                                "PAST.MEDICAL.HISTORY.CODE
",
                                                "MODE.OF.ARRIVAL",
                                                "STATE.AT.THE.TIME.OF.ARRI
VAL",
                                                "TYPE.OF.ADMSN",
                                                "TOTAL.COST.TO.HOSPITAL"
                                                #"Log.Cost.Treatment"
)])
```

---

## Model building: Using the lm() function

The actual model building starts now. Note that we are demonstrating the strategy of building a step wise model (forward selection and backward elimination) using the **lm()** function

```
# Null Model
noModel <- lm(TOTAL.COST.TO.HOSPITAL ~ 1, data = reg.train.data)

# Full Model
RegModelFull = lm(TOTAL.COST.TO.HOSPITAL ~ ., data = reg.train.data)
```

```
# Stepwise - Forward selection backward elimination
RegModelStepwise <- step(noModel, list(lower = formula(noModel), upper
= formula(RegModelFull)),
    direction = "both", trace = 0)
```

---

## Model Evaluation

### 1. Model summary of Train Data

Checking the if the model satisfies the assumpations of Linear Regression Model. Note that this evaluation is on training data.

The model summary gives the equation of the model as well as helps test the assumption that beta coeffiecents are not statically zero.

```
summary(RegModelStepwise)

##
## Call:
## lm(formula = TOTAL.COST.TO.HOSPITAL ~ TYPE.OF.ADMSN + I_COST.OF.IMPL
ANT +
##     AGE + HR.PULSE, data = reg.train.data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -173132  -54897   -4510   35478  374285
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)             4.219e+04  4.709e+04   0.896   0.3720
## TYPE.OF.ADMSNEMERGENCY 1.326e+05  2.551e+04   5.195 7.99e-07 ***
## I_COST.OF.IMPLANT       2.222e+00  3.835e-01   5.794 5.18e-08 ***
## AGE                     1.852e+03  3.830e+02   4.836 3.78e-06 ***
## HR.PULSE                8.028e+02  4.413e+02   1.819   0.0713 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 85690 on 126 degrees of freedom
## Multiple R-squared:  0.5635, Adjusted R-squared:  0.5497
## F-statistic: 40.67 on 4 and 126 DF,  p-value: < 2.2e-16
```

You may ignore the below code chunk. This code is more to understand how the standard error of beta coefficients are calculated. **vcov()** is used to compute the variance covariance matrix of the fitted object. **cov2cor()** is used to scale the covariance matrix into the correponding correlation matrix. In the matrix generated out of the below code chunk:

1. The diagonal values are the variance of the variable to itself.
2. The square root of the diagonal values gives the standard error associated with the estimates.
3. The non diagonal elements are the covariance of the estimated

```
vcov(RegModelStepwise)
```

```
##                           (Intercept) TYPE.OF.ADMSNEMERGENCY
## (Intercept)              2.217524e+09           2.992403e+08
## TYPE.OF.ADMSNEMERGENCY   2.992403e+08           6.509817e+08
## I_COST.OF.IMPLANT       -6.980151e+02           6.736939e+00
## AGE                     -1.158026e+07          -5.031169e+06
## HR.PULSE                -2.011840e+07          -2.538171e+06
##                         I_COST.OF.IMPLANT            AGE      HR.PULSE
## (Intercept)                -698.0150674 -1.158026e+07 -2.011840e+07
## TYPE.OF.ADMSNEMERGENCY        6.7369385 -5.031169e+06 -2.538171e+06
## I_COST.OF.IMPLANT             0.1470902 -2.184030e+01  6.417772e-01
## AGE                         -21.8403041  1.466900e+05  8.643503e+04
## HR.PULSE                      0.6417772  8.643503e+04  1.947850e+05
```

```
sqrt(diag(vcov(RegModelStepwise)))
```

```
##             (Intercept) TYPE.OF.ADMSNEMERGENCY      I_COST.OF.IMPLANT
##            4.709060e+04           2.551434e+04           3.835234e-01
##                     AGE               HR.PULSE
##            3.830013e+02           4.413445e+02
```

```
#cov2cor(vcov(RegModelStepwise))
```

## 2. The residual analysis

The error term diagnostic is critical to understanding the behaviour of linear regression models. The two critical assumptions of linear regression are:

1. Error term should be normally distributed
2. Error term should have constant variance (**homoscedasticity**)

The **plot()** function when used on the regression object model gives us four different plots. The two important one to analyze there are:

1. Normal Q-Q
2. Scale-Location

### *1. Normal Q_Q plot*

This plot shows if the error terms are normally distributed. In case, of normal distribution, the dots should appear close to the straight line with not much of a deviation.

### *2. Scale-Location*

Also known as spread location plot, it shows if the residuals are equally spread along the range of predictors. It is desirable to see a horizontal straight line with with randomly spread points.

**The other two plots are:**

### *3. Residual vs. Fitted*

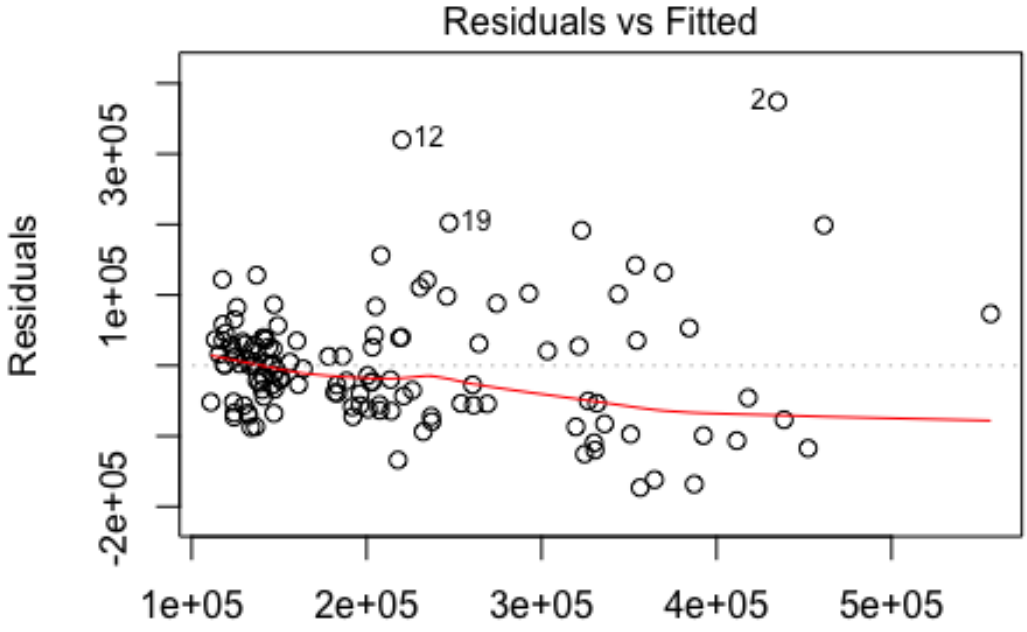
There could be a non linear relationship between predictor variable (Xs) and the outcome variable (Y). This non linear relationship can show up in this plot which may suggest that the model is mis-specified. It is desirable to see a horizontal straight line with with randomly spread points.

### *4. Residual vs. Leverage*

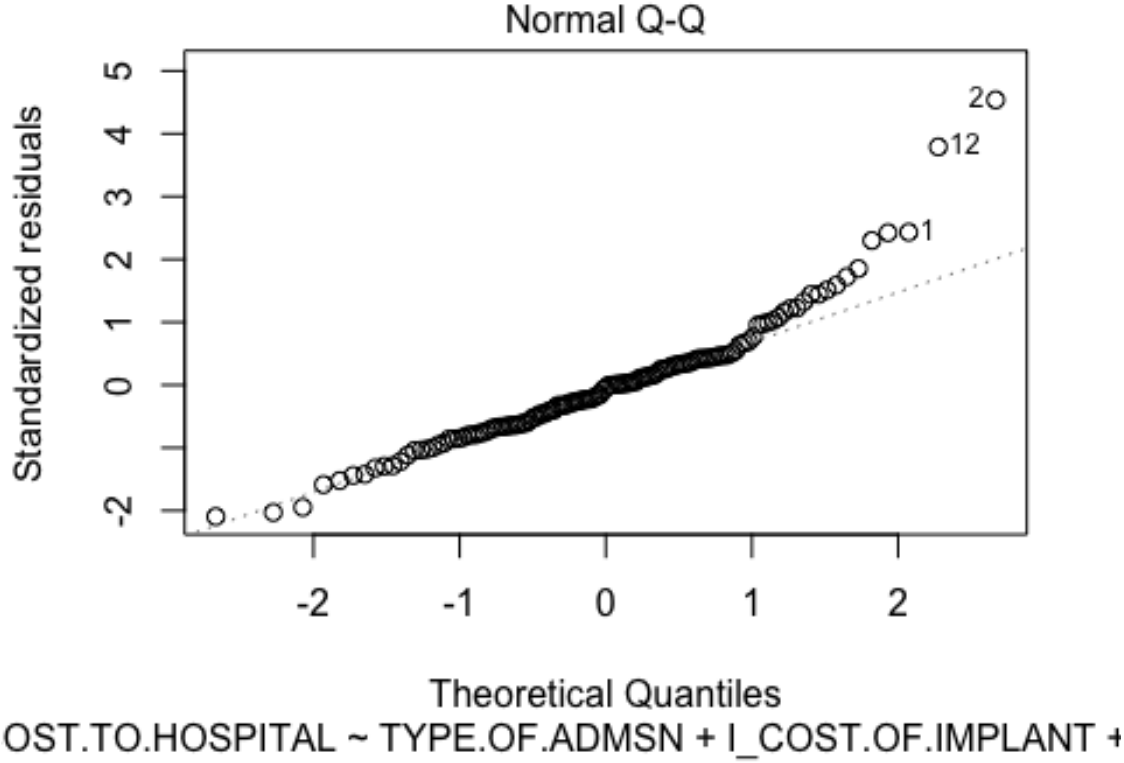
The regression line can be influenced by outliers (extreme values in Y) or by data points with high leverage (extreme values in X). Not all the extreme values are influential cases in regression analysis.

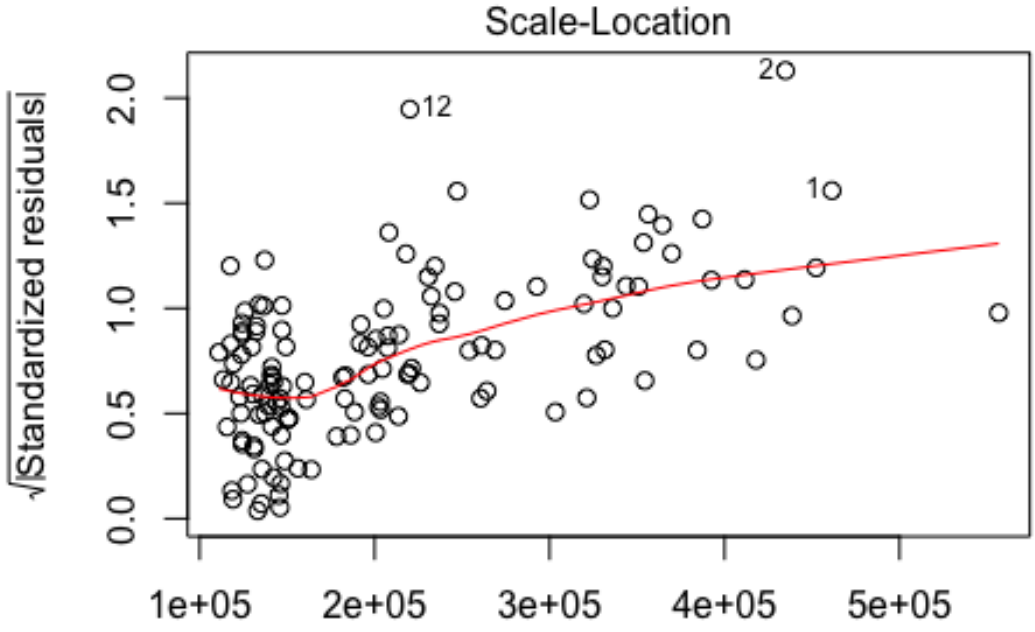
Even if data has extreme values, it may not be influential to determine the regression line. On the flip side, some cases could be very influential even if they do not seem to be an outlier. Influential cases are identified by cook's distance. In the plot, look for for outlying values at the upper right corner or at the lower right corner (cases outside of a dashed line i.e. Cook's distance).

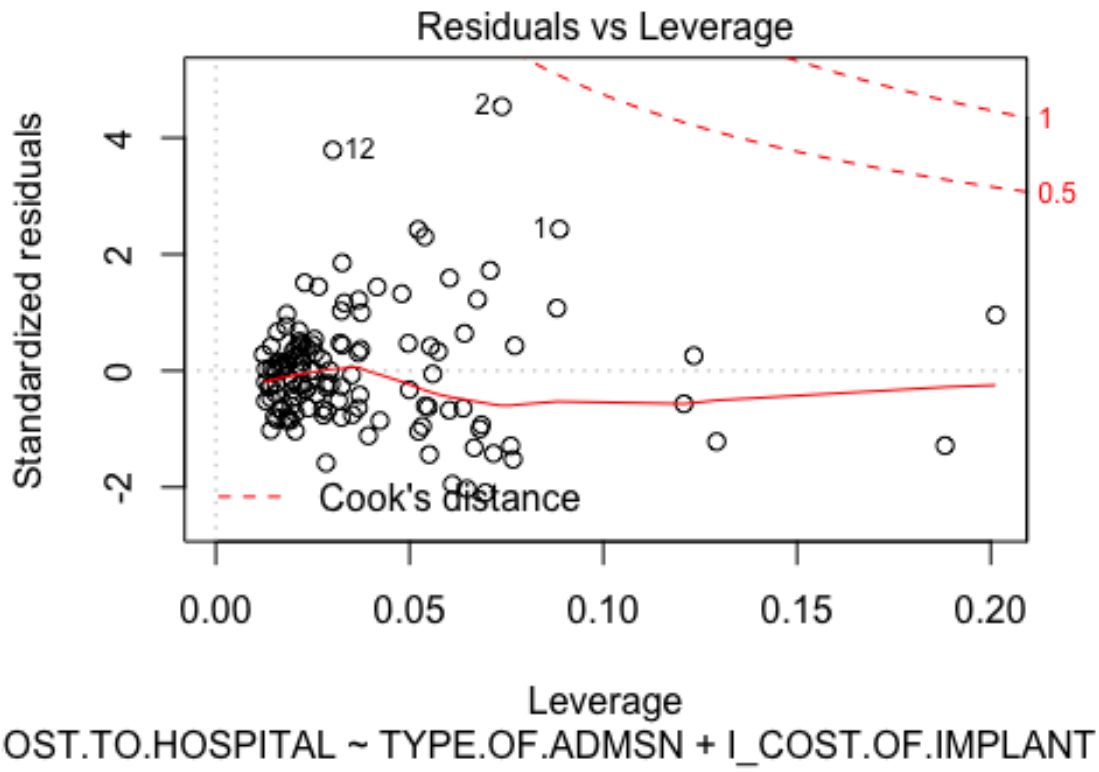
```
plot(RegModelStepwise)
```




OST.TO.HOSPITAL ~ TYPE.OF.ADMSN + I_COST.OF.IMPLANT +


Normal Q-Q
Standardized residuals
5
4
3
2
1
0
-2
2
12
1
-2
-1
0
1
2
Theoretical Quantiles
OST.TO.HOSPITAL ~ TYPE.OF.ADMSN + I_COST.OF.IMPLANT +


Scale-Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
12
2
1
1e+05
2e+05
3e+05
4e+05
5e+05
Fitted values
OST.TO.HOSPITAL ~ TYPE.OF.ADMSN + I_COST.OF.IMPLANT +

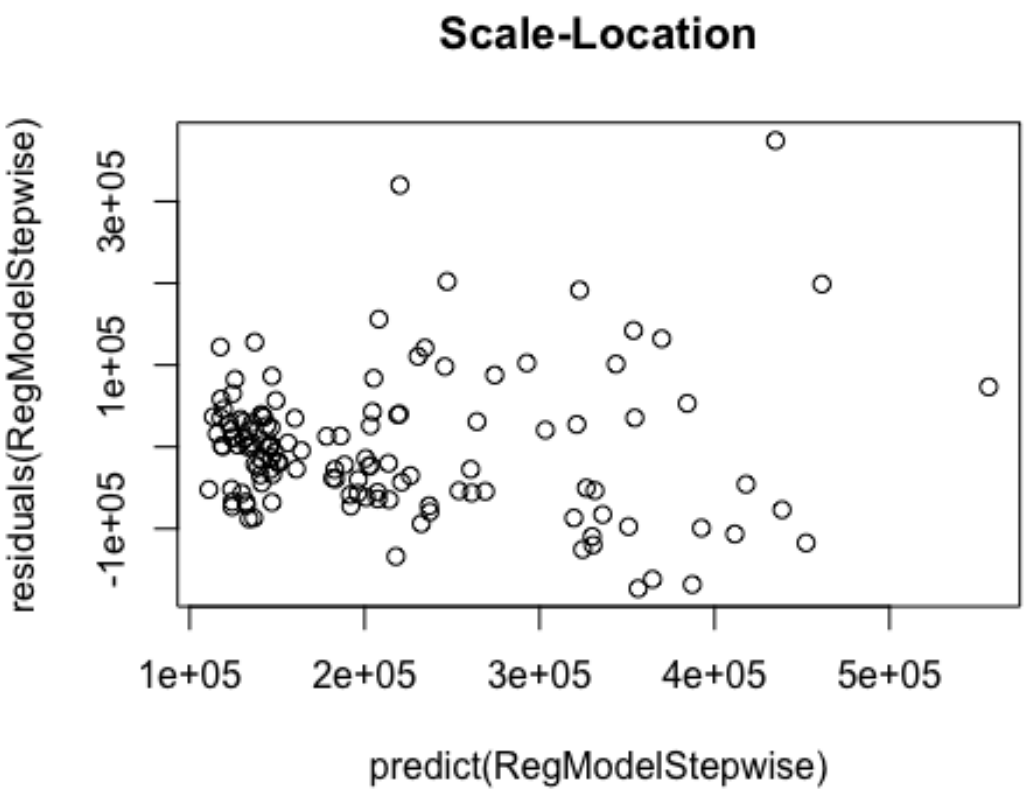



OST.TO.HOSPITAL ~ TYPE.OF.ADMSN + I_COST.OF.IMPLANT +

```
#hist(residuals(RegModelStepwise), main = "Residuals", col = 'blue')
```

*Visual inspection to check for heteroscedasticity in error terms*

You may ignore the below code chuck. This is an elaboration of the scale-location plot obtained before.




*Multi-colinearity*

Variance Inflation Factor (VIF) is a measure of how much the variance of the estimated regrression coeffiecients are inflated as compared to when the predicator variable are not linearly related.

VIF = 1 : Not Correlated 1<VIF<5 : Moderately Correlated 5<VIF<=10: Highly Correlated

*The square root of the VIF tells you how much larger the standard error is, compared with what it would be if that variable were uncorrelated with the other predictor variables in the model.*

Say, if the square root of the VIF is 2.5; this means that the standard error for the coefficient of that predictor variable is 2.5 times as large as it would be if the predictor variable were uncorrelated with the other predictor variables

Generally the issue of multi-colinearity wil not arise, if the corelation amongst variable has been analyzed before model building and the one amongst the corelated variable has been dropped from the data.

```
vif(RegModelStepwise)
```

```
##     TYPE.OF.ADMSN I_COST.OF.IMPLANT                 AGE          HR.PU
LSE
```

```
##          1.376674          1.041076          1.825443          1.371
082
```

## 3. Model Validation on the Test Data

The **predict** function is used to get the predicted response on the new dataset. You may get an error message if the test data has got any new levels which was not there in the training set. This generally happens when the data has categorical variable with multiple levels.

```
RegTestPrediction = predict(RegModelStepwise, reg.test.data, interval =
"confidence",
    level = 0.95, type = "response")
print(RegTestPrediction)
```

```
##          fit       lwr      upr
## 7   264104.8 215718.08 312491.4
## 13  634002.6 490645.39 777359.9
## 20  193467.0 170843.51 216090.6
## 24  369427.5 327542.44 411312.6
## 33  433188.5 387575.62 478801.4
## 37  109551.7  80166.53 138936.9
## 53  231500.1 196711.87 266288.3
## 86  228070.5 191182.38 264958.6
## 88  314826.3 266919.59 362733.1
## 89  129623.4 101891.59 157355.2
## 94  444822.5 401178.34 488466.8
## 96  226436.7 192016.58 260856.8
## 110 133084.7 111288.03 154881.3
## 124 179253.1 154442.74 204063.5
## 136 141606.0 119147.70 164064.2
## 142 135552.1 113888.45 157215.8
## 145 137282.8 117091.95 157473.6
## 149 125550.4 102673.03 148427.7
## 157 145929.1 118889.16 172969.1
## 159 393695.1 348515.76 438874.4
## 166 322994.8 275200.84 370788.9
## 170 226436.7 192016.58 260856.8
## 194 352447.5 309766.81 395128.2
## 199 156302.7 125796.36 186809.1
## 202 130367.1 109069.29 151664.8
## 203 138641.6 117828.89 159454.3
## 213 119993.5  95556.20 144430.7
## 223 134686.8 109920.51 159453.1
## 224 191380.4 163541.37 219219.5
## 228 363377.1 321956.64 404797.6
## 237 228689.1 194313.69 263064.5
## 248 260487.5 230593.46 290381.6
```

## End of Document