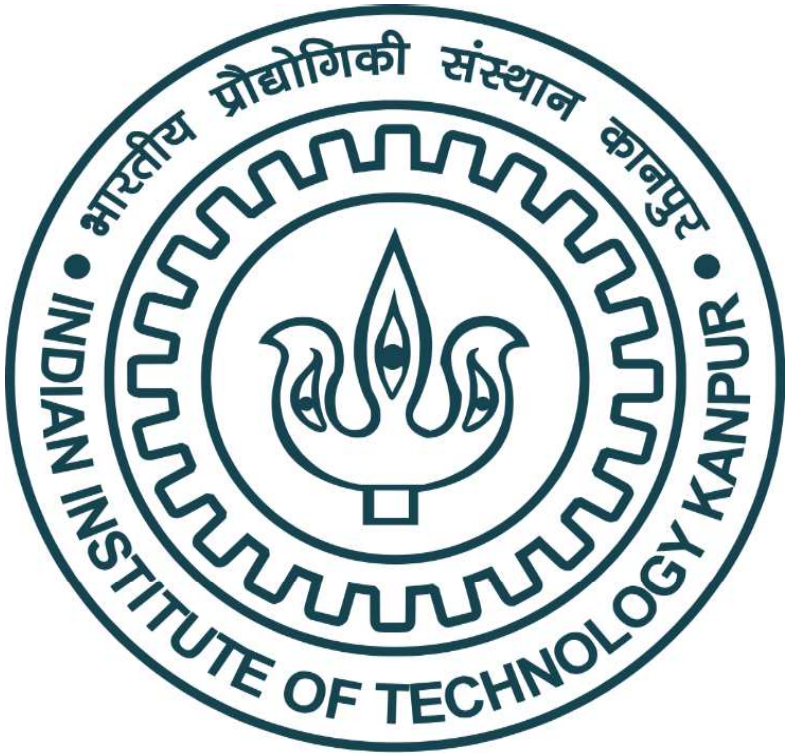


A regression Analysis Project for Predicting future Medical Expenses of Individuals

BY:

VIJAY KUMAR (201454)

SAHIL YADAV (201399)

RAVINDRA CHAURASIYA (201383)

HARSHIT GARG (201321)

PAWAN KUMAR (201363)

Acknowledgement

We would like to thank our instructor Dr.Sharmistha Mitra Department of Mathematics and Statistics, IIT Kanpur for her guidance, monitoring and constant encouragement for this project without which it is not possible task to do.

We would like to thank her for giving us this opportunity to explore the real life application of regression models and hence giving us an idea about their statistical viewpoint in day to day life.

Introduction

The purposes of this exercise to look into different features to observe their relationship, and plot a multiple linear regression based on several features of individual such as age, physical/family condition and location against their existing medical expense to be used for predicting future medical expenses of individuals that help medical insurance to make decision on charging the premium.

Description Of Data

We have 6 regressors out of which 2 are binary categorical regressors (sex and smoker)and 1 is a categorical regressor (region) with 4 categories. We introduce 3 dummy variable for region regressor(X_6,X_7,X_8) and 1 dummy variable each for sex and smoking. We have a total of 1338 data points.

Y: Expenses

X_1**:** Age

X_2**:** Sex

X_3**:** BMI

X_4**:** Children

X_5**:** Smoker

X_6**:** SE(South East)

X_7**:** NW(North West)

X_8**:** NE(North East)

Model and Assumptions

Model: The linear regression model is given by:

$y_i = \beta_0 \sum_{j=1}^{8} x_{ij}\beta_j + \epsilon_i \quad , i = 1\,(1)\,1338$

Assumptions:

$(i)\,\epsilon_i's$ are normally distributed.

$(ii)\,E\,(\epsilon_i) = 0, \forall i = 1(1)n \;\; and \;\; Var\,(\epsilon_i) = \sigma^2, \forall i = 1(1)n$ i.e the errors are homoscedastic.

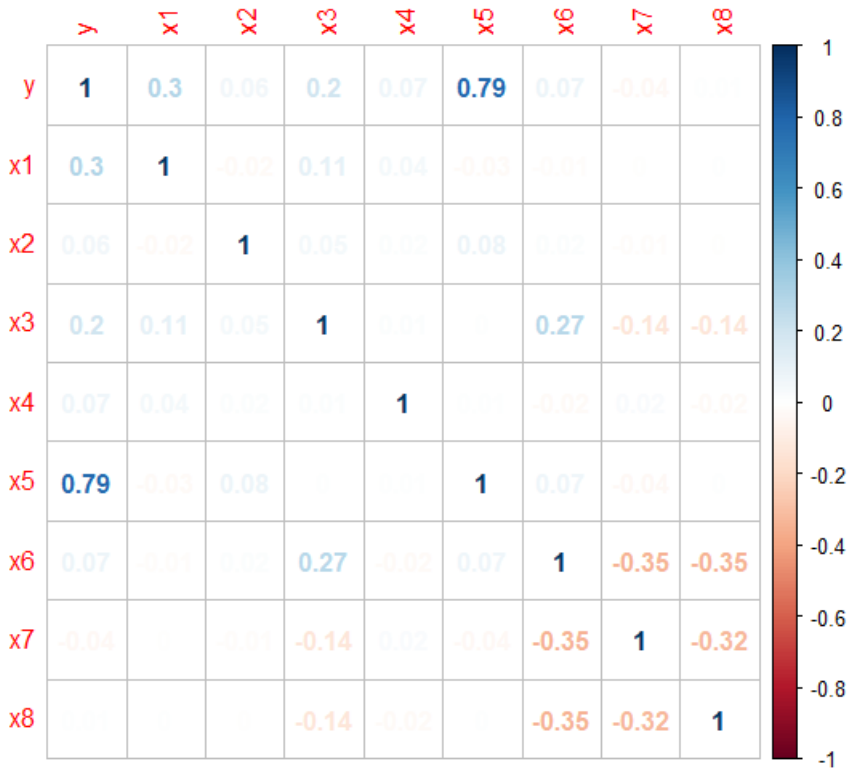
$(iii)\,(\epsilon_1, \epsilon_2, \ldots, \epsilon_n)$ are independentely distributed i.e $\text{Cov}(\epsilon_i, \epsilon_k) = 0, \forall i \neq k$ i.e errors are uncorrelated.

$(iv)\,Cov\,(x_i, x_k) = 0 \;\; \forall i \neq k$ i.e the model is free from multicollinearity problem.

Statistical Analysis

Initial Model:
$\mathbf{Y}=\beta_0 + \beta_1\ X_1 + \beta_2\ X_2 + \beta_3\ X_3 + \beta_4\ X_4 + \beta_5\ X_5 + \beta_6\ X_6 + \beta_7\ X_7 + \beta_8\ X_8 + \epsilon$

Plot of correlation coefficient of response variable (Y) and regressors (X_j ; j= 1(1) 8) is shown below:



Positive and Negative correlation is indicated by color Blue and Red respectively. From the correlogram we can see that only X_5 is correlated with response variable (Y)

Summary of Full Model

```
Call:
lm(formula = y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8, data = Insurancefinal_)

Residuals:
     Min       1Q   Median       3Q      Max
-11302.7  -2850.9   -979.6   1383.9  29981.7

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -12900.87    1020.90 -12.637  < 2e-16 ***
x1             256.84      11.90  21.586  < 2e-16 ***
x2            -131.35     332.94  -0.395 0.693255
x3             339.29      28.60  11.864  < 2e-16 ***
x4             475.69     137.80   3.452 0.000574 ***
x5           23847.48     413.14  57.723  < 2e-16 ***
x6             -76.29     470.64  -0.162 0.871253
x7             606.52     477.18   1.271 0.203940
x8             959.31     477.91   2.007 0.044921 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6062 on 1329 degrees of freedom
Multiple R-squared:  0.7509,    Adjusted R-squared:  0.7494
F-statistic: 500.9 on 8 and 1329 DF,  p-value: < 2.2e-16
```

Multicollinearity:

Multicollinearity refers to a situation in which more than two explanatory variables in a multiple regression model are highly linearly related. This problem can led to a huge change in the estimated on the addition/removal of a data point. There are various ways to check whether multicollinearity exists or not. We used two methods to check whether multicollinearity exists or not.

(i) **Variance Inflation Factor**

VIF is calculated for parameters corresponding to each regressor. Intuitively it denotes a factor with which the variance of the estimated parameters inflates. VIF is calculated as:

$$VIF_j = \frac{1}{1 - R_j^2} \quad \forall j = 1 \dots p$$

where R_j^2 = Coefficient of determination when x_j is regressed with other regressors. The calculated VIF for all the parameters are:

x1	x2	x3	x4	x5	x6	x7	x8
1.016843	1.008900	1.106682	1.004008	1.012067	1.597204	1.524717	1.526170

Since all VIF's are less than 5 therefore regressors are free from Multicollinearity.

(ii) **Variance Decomposition Method**

It is a method to identify subsets of regressors involved in multicollinearity. The Variance decomposition matrix (Π) is calculated as:

$$X = UDV^T$$

$$(\ \Pi)_{\ kj} = \frac{\frac{1}{\lambda_k} v_{kj}^2}{\sum_{k=1}^{p} \frac{1}{\lambda_k} v_{kj}^2} \quad \forall k, j = 1 \dots p$$

The Calculated matrix is shown below:

	Eigenvalues	CI	(Intercept)	x1	x2	x3	x4	x5	x6	x7	x8
1	5.0021	1.0000	0.0010	0.0040	0.0118	0.0013	0.0119	0.0090	0.0055	0.0049	0.0049
2	1.0143	2.2207	0.0000	0.0001	0.0000	0.0000	0.0036	0.0292	0.2138	0.2296	0.0042
3	1.0006	2.2358	0.0000	0.0000	0.0000	0.0000	0.0002	0.0000	0.0520	0.1002	0.3389
4	0.7491	2.5840	0.0003	0.0018	0.0003	0.0005	0.0222	0.9327	0.0230	0.0070	0.0008
5	0.5153	3.1156	0.0002	0.0006	0.2416	0.0003	0.7421	0.0231	0.0026	0.0114	0.0032
6	0.4295	3.4127	0.0019	0.0135	0.7009	0.0026	0.1771	0.0000	0.0470	0.0537	0.0472
7	0.1999	5.0020	0.0053	0.1205	0.0185	0.0107	0.0330	0.0002	0.5727	0.5074	0.5155
8	0.0727	8.2938	0.0506	0.8217	0.0233	0.1412	0.0052	0.0025	0.0835	0.0330	0.0320
9	0.0164	17.4427	0.9408	0.0378	0.0034	0.8434	0.0047	0.0032	0.0000	0.0528	0.0533

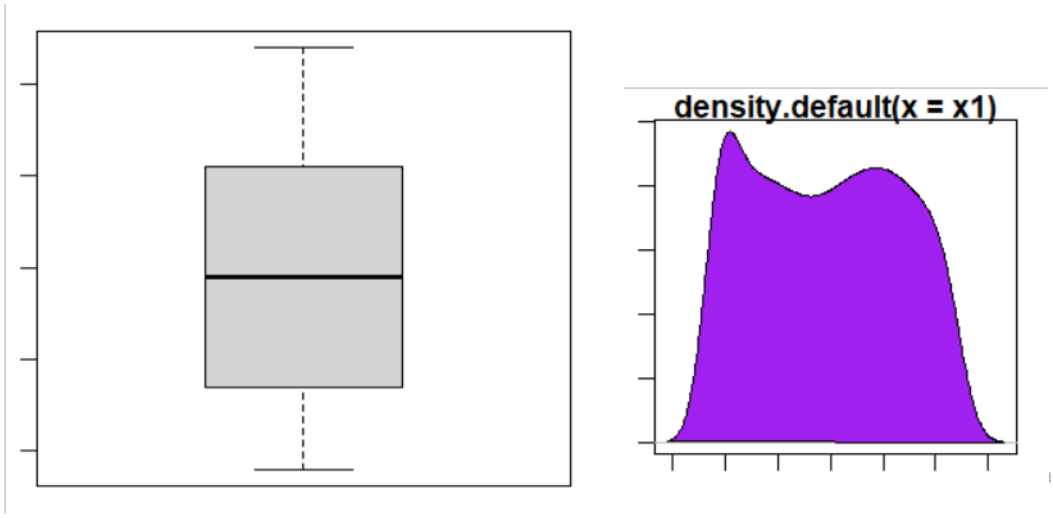
Since maximum condition index is 17.4427 which is > 15 but there are no such corresponding subset of regressors which are > 0.5 hence our model is free from Multicollinearity.

Checking Skewness of Regressors

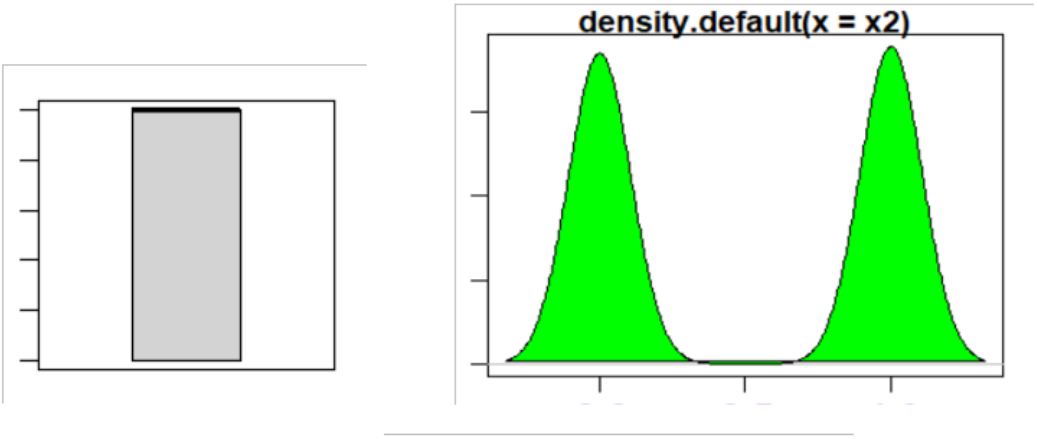
```
> skewness(x1)
[1] 0.05561008
> skewness(x2)
[1] -0.0209279
> skewness(x3)
[1] 0.2842738
> skewness(x4)
[1] 0.9373281
> skewness(x5)
[1] 1.463124
> skewness(x6)
[1] 1.024471
> skewness(x7)
[1] 1.199063
> skewness(x8)
[1] 1.203809
```

As we could see values of skewness fall between - 3 and + 3 for all regressors which is indication of acceptable skewness, hence we have no outliers in data. Box plot of regressors and there corresponding densities are :

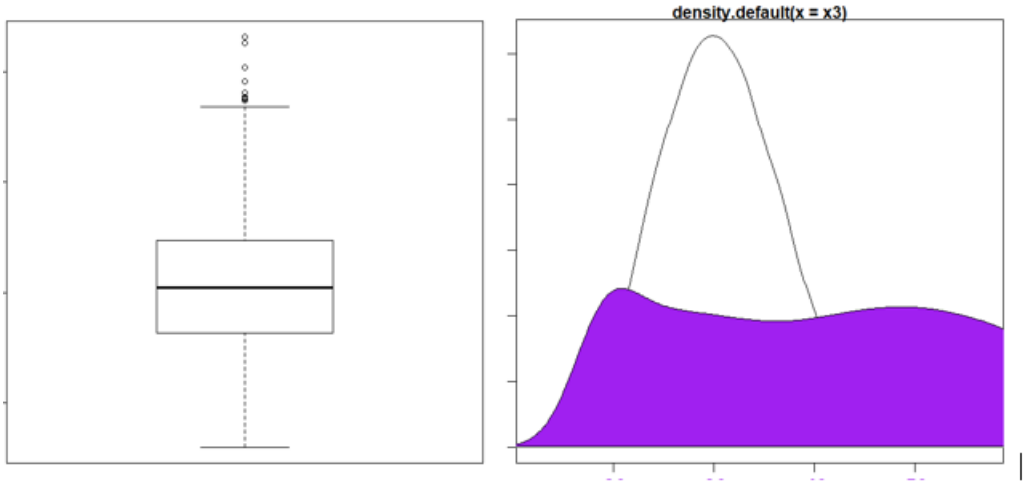
(i) X_1



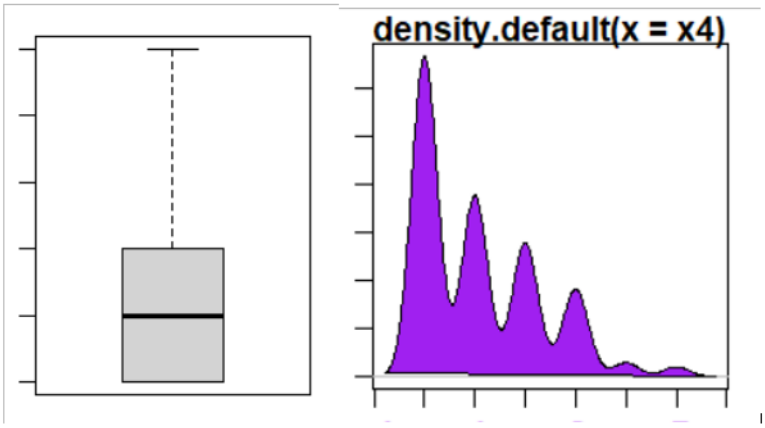
(ii) X_2



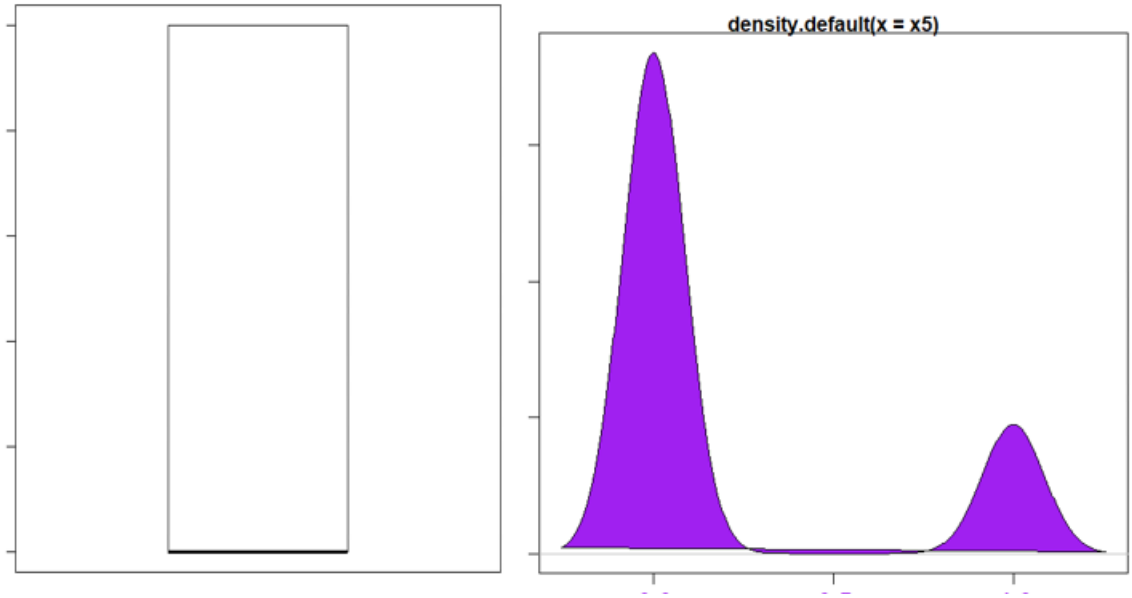
(ii) X_3



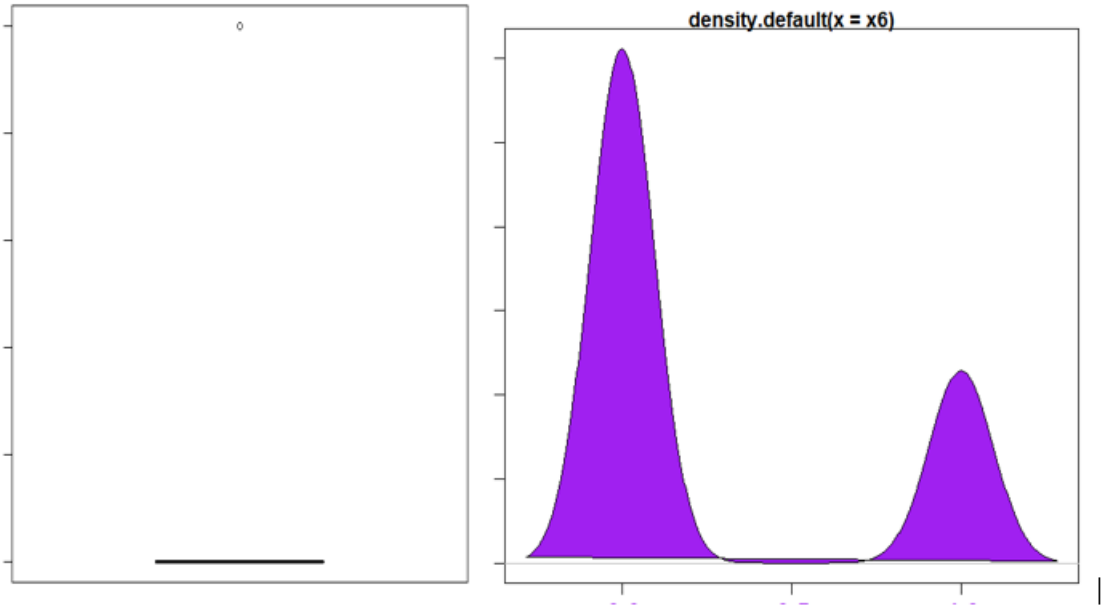
(ii) X_4



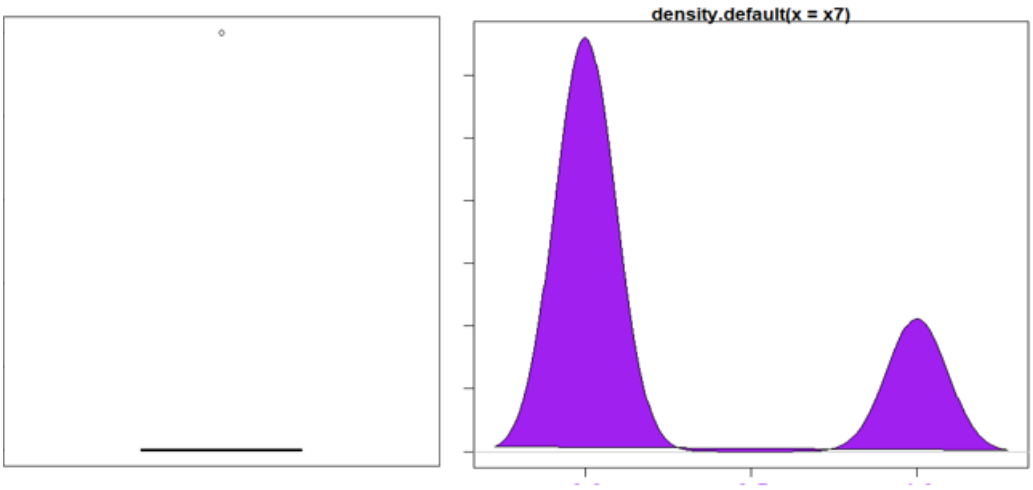
(ii) X_5



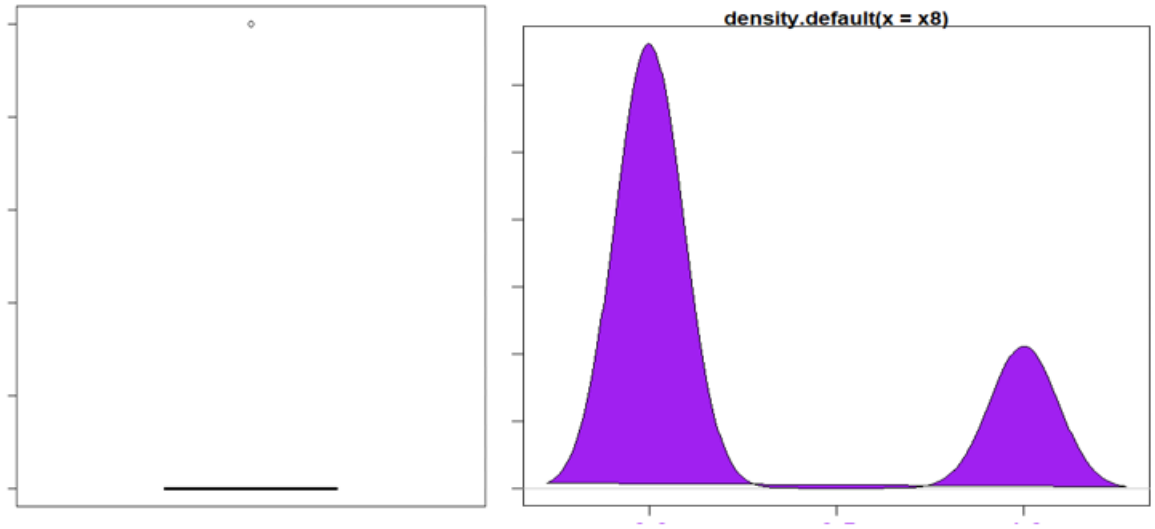
(ii) X_6



(ii) X_7



(ii) X_8

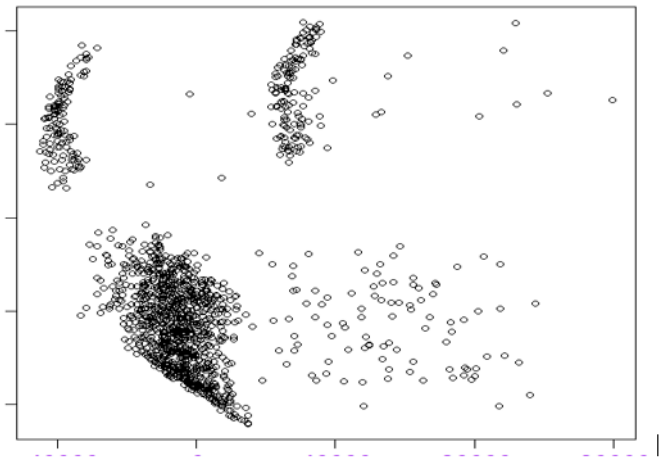


Residual Analysis

Residual analysis is very important so as to check if our error assumptions on the OLS model are indeed true and if not, taking suitable steps to modify the model. We do this by plotting graphs between the residuals and different regressors to see if there is some apparent relation between them. Note that this method can check for heteroscedasticity in the model.

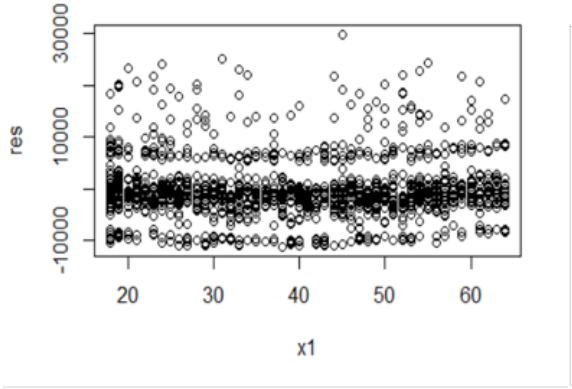
Residual Plots:

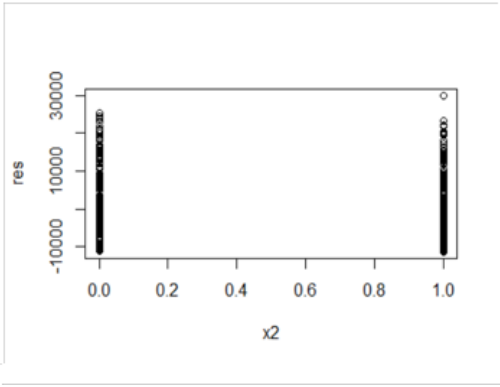
1. Residuals ($\hat{e}_i$'s) are plotted against predicted responses ($\hat{y}_i$'s)

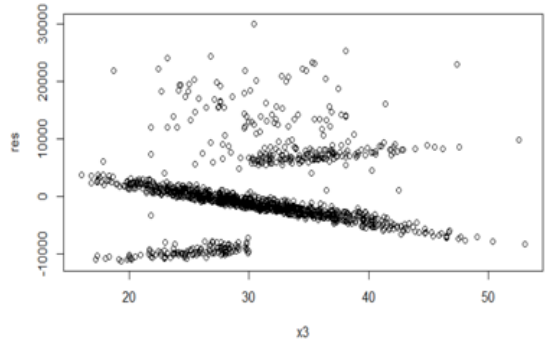


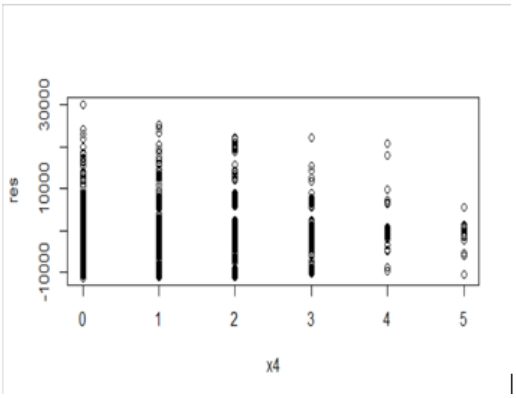
Now according to theory if our error assumptions were true we must have got totally random plot. However in the above residual plot there is an evident pattern. This means that the errors are correlated with predicted responses which is the linear combination of regressors.
To find out which regressor is causing the issue we need to plot the residuals against each regressor.

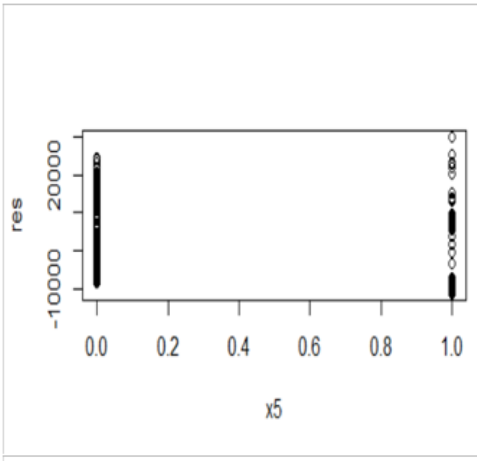
2. Residuals ($\hat{e}_i$'s) are plotted against each Regressor

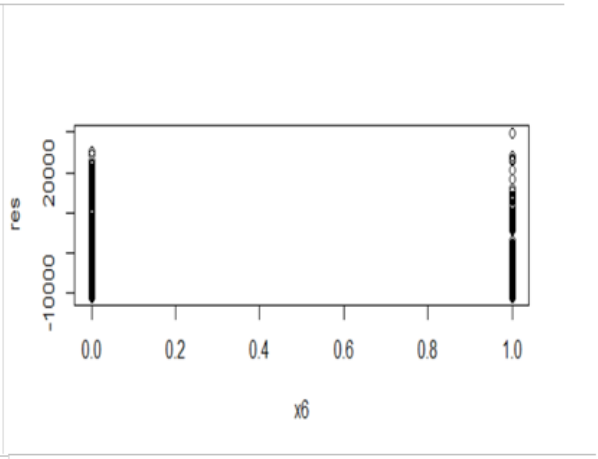


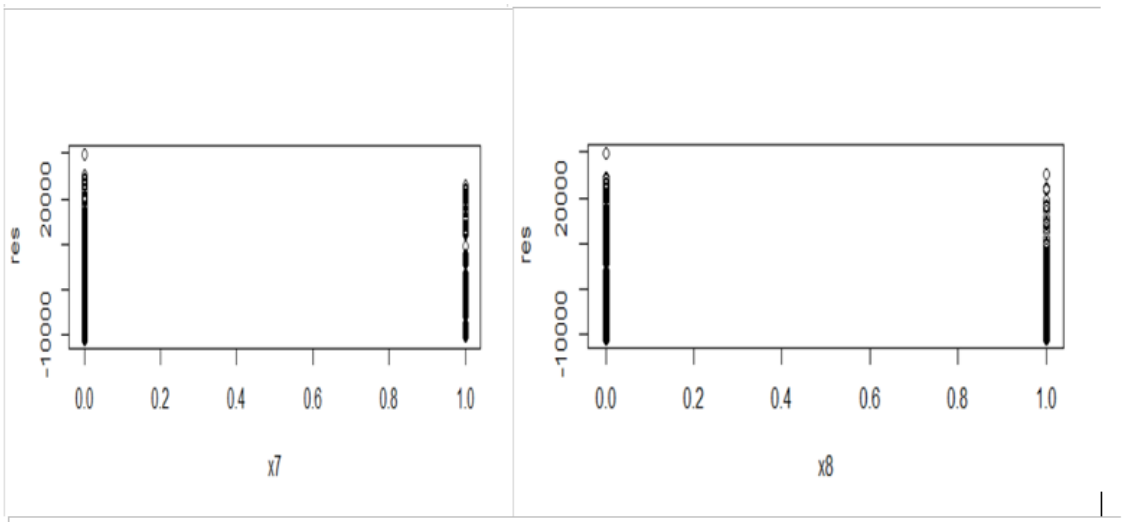












So from the above plots we can clearly see that all the plots are totally random except X_3(BMI) regressor. The residualss plot against X_3 is close to an outward opening funnel which suggest the following relationship:

$$Var(e_i) \propto \sigma^2 BMI_i^2$$

So now to ensure our usual error assumptions, we need to make changes to our model. The way we can get the above relationship is by dividing the model on both sides by the X_3(BMI) regressor. So our new model is:

$y_i^* = \sum_{j=1}^{p} \beta_j^* x_{ij}^* + e_i^*$ where

$y_i^* = \frac{y_i}{BMI}$ $x_{ij}^* = \frac{x_{ij}}{BMI}$ $e_i^* = \frac{e_i}{BMI}$

Note that $Var(e_i^*) = \frac{Var(e_i)}{BMI^2} \propto \sigma^2$ and thus our usual assumptions are satisfied now. To see if we have actually removed heteroscedasticity problem after considering the above transformation. There are few methods which we will consider.

1. **Breush - Pagan test:**
the Breusch–Pagan test is used to test for heteroscedasticity in a linear regression model. It tests whether the estimated variance of the residuals from a regression model are dependent on the values of the independent variables. In that case, we have heteroscedasticity in our model.

Testing Criteria:

H_0: Homoscedasticity is present.

H_A: Heteroscedasticity is present.

If the p-value of the test is less than some significance level (i.e. $\alpha = .05$) then we reject the null hypothesis and conclude that heteroscedasticity is present in the regression model.

So after applying Breusch–Pagan test on different models i.e Original Model,WLS Model, Log transformation Model we got following results:

(i) Original Model

```
        studentized Breusch-Pagan test

data:  fitt
BP = 121.59, df = 8, p-value < 2.2e-16
```

(ii) WLS Model

```
        studentized Breusch-Pagan test

data:  fitt11
BP = 121.59, df = 8, p-value < 2.2e-16
```

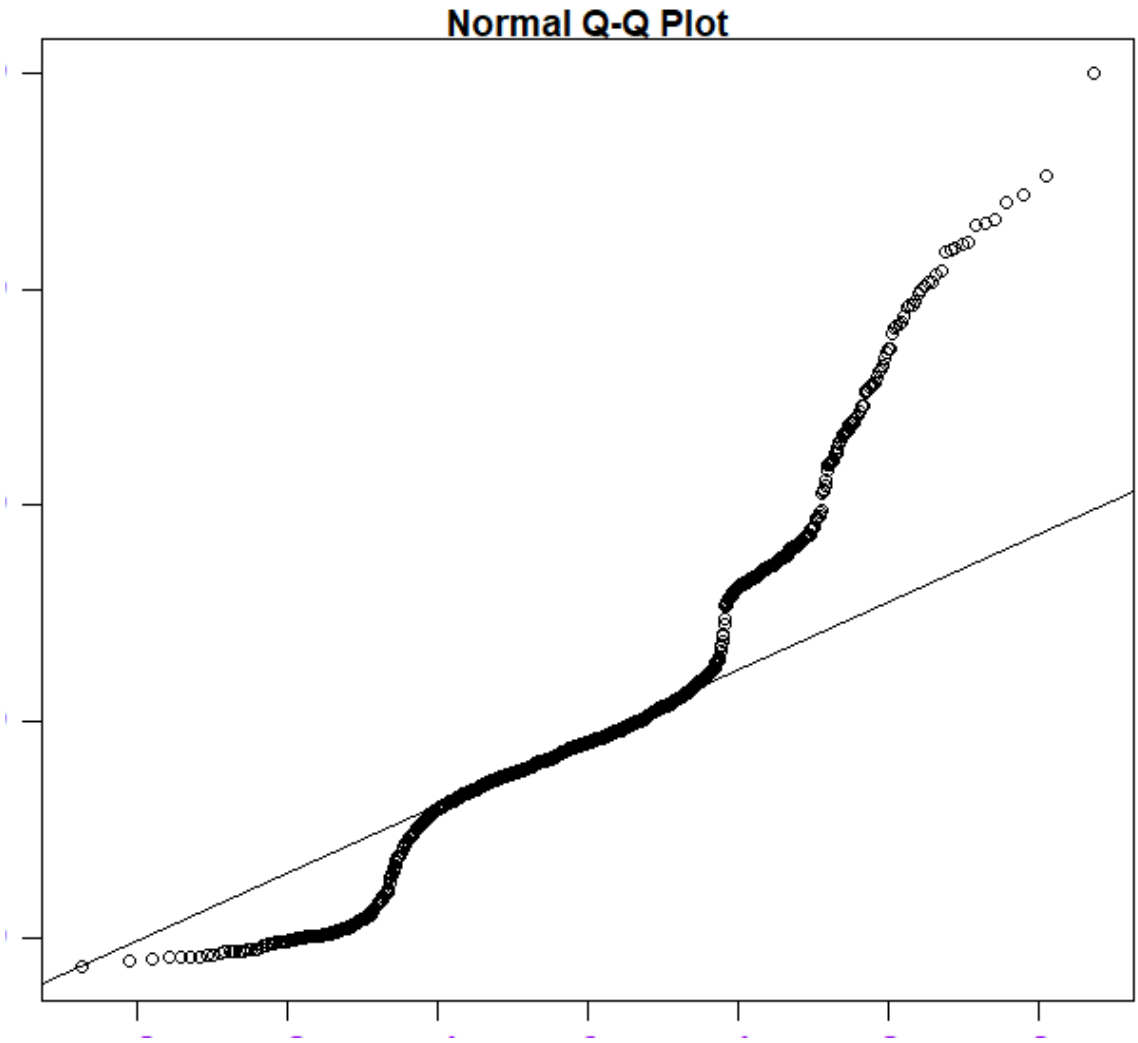
(iii) Log Transformation Model

```
        studentized Breusch-Pagan test

data:  fitt2
BP = 77.071, df = 8, p-value = 1.897e-13
```

So from the above results it can be seen that out of the 3 models only Log Transformation Model was able to eradicate Heteroscedasticity to some extent as it has the highest p - value out of the 3 models.

Checking Normality



Shapiro - Wilk Test:

The Shapiro–Wilk test is a test of normality of a random sample.

H_0:Sample comes from normally distributed population

H_A: Sample does not come from normal population

Test Statistic

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_i\right)^2}{\sum_{i=1}^{n} \left(x_i - \overline{x}\right)^2}$$

where, x_i's are ordered random sample
a_i's are the constant generated from the covariances, variances and means of the sample (size n) from a normally distributed sample.

Decision Rule:

We reject H_0 if $W < W_\alpha$ at $\alpha\%$ level of significance otherwise accept H_0

```
        Shapiro-Wilk normality test

data:  res
W = 0.89895, p-value < 2.2e-16
```

Since p value < 0.05 we reject Null Hypothesis and population is not normally distributed.

Variable Selection Method

Building a regression model that includes only a subset of the available regressors involves two conflicting factors:

(i) Include as many regressors as possible so that the information content in these factors can influence the predicted value of y.
(ii) Include as few regressors as possible because the variance of the prediction $\hat{y}$ increases as the number of regressor increase.

The process of finding a model that is a compromise between these two objectives is called the best regression equation.

Akaike Information Criteria:

AIC ia measure of the relative quality of statistical models for a given set of data.Given a colllection of models for the data, AIC estimates the quality of each model, relative to each of the other models.Hence AIC provides a means for model selection.

AIC= $n \ln SSRes_(p) + 2p$

Compute AIC for all possible models and choose the model for which AIC is minimum.

The output is given below:

```
Start:  AIC=23316.34
y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8

       Df  Sum of Sq        RSS   AIC
- x6    1 9.6556e+05 4.8837e+10 23314
- x2    1 5.7197e+06 4.8842e+10 23315
- x7    1 5.9365e+07 4.8896e+10 23316
<none>               4.8837e+10 23316
- x8    1 1.4806e+08 4.8985e+10 23318
- x4    1 4.3789e+08 4.9274e+10 23326
- x3    1 5.1723e+09 5.4009e+10 23449
- x1    1 1.7122e+10 6.5958e+10 23717
- x5    1 1.2244e+11 1.7127e+11 24993

Step:  AIC=23314.37
y ~ x1 + x2 + x3 + x4 + x5 + x7 + x8

       Df  Sum of Sq        RSS   AIC
- x2    1 5.7200e+06 4.8843e+10 23313
<none>               4.8837e+10 23314
- x7    1 8.8108e+07 4.8926e+10 23315
+ x6    1 9.6556e+05 4.8837e+10 23316
- x8    1 2.1080e+08 4.9048e+10 23318
- x4    1 4.3949e+08 4.9277e+10 23324
- x3    1 5.3016e+09 5.4139e+10 23450
- x1    1 1.7145e+10 6.5982e+10 23715
- x5    1 1.2290e+11 1.7174e+11 24995

Step:  AIC=23312.53
y ~ x1 + x3 + x4 + x5 + x7 + x8

       Df  Sum of Sq        RSS   AIC
<none>               4.8843e+10 23313
- x7    1 8.8139e+07 4.8931e+10 23313
+ x2    1 5.7200e+06 4.8837e+10 23314
+ x6    1 9.6588e+05 4.8842e+10 23315
- x8    1 2.1052e+08 4.9054e+10 23316
- x4    1 4.3789e+08 4.9281e+10 23323
- x3    1 5.2970e+09 5.4140e+10 23448
- x1    1 1.7171e+10 6.6014e+10 23714
- x5    1 1.2347e+11 1.7232e+11 24997

Call:
lm(formula = y ~ x1 + x3 + x4 + x5 + x7 + x8)

Coefficients:
(Intercept)           x1           x3           x4           x5           x7           x8
   -12969.0        257.0        338.0        475.4      23830.9        644.5        996.7
```

So from above model 3 has the minimum AIC value i.e 23312.53.

The final model is y= $\beta_0 + \beta_1 x_1 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_7 x_7 + \beta_8 x_8 + \epsilon$

So now we used Box cox method to normalize the data. So our transformed model is:

$\frac{y^\lambda - 1}{\lambda} = \beta_0 + \beta_1 x_1 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_7 x_7 + \beta_8 x_8 + \epsilon$ where λ= 0.1414141

Transformed Model Characteristics:

```
Call:
lm(formula = ((y^lambda - 1)/lambda) ~ x1 + x3 + x4 + x5 + x7 +
    x8)

Residuals:
    Min      1Q  Median      3Q     Max
-3.5691 -0.7675 -0.2533  0.1851  7.7469

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 10.721493   0.263980  40.615  < 2e-16 ***
x1           0.118657   0.003116  38.076  < 2e-16 ***
x3           0.053280   0.007378   7.221 8.63e-13 ***
x4           0.326558   0.036097   9.047  < 2e-16 ***
x5           5.870846   0.107754  54.484  < 2e-16 ***
x7           0.266831   0.109071   2.446   0.0146 *
x8           0.480759   0.109139   4.405 1.14e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.589 on 1331 degrees of freedom
Multiple R-squared:  0.7749,    Adjusted R-squared:  0.7739
F-statistic: 763.7 on 6 and 1331 DF,  p-value: < 2.2e-16
```

Now applying Breusch Pagan Test to check Heteroscedasticity

```
        studentized Breusch-Pagan test

data:  new_model
BP = 55.176, df = 6, p-value = 4.271e-10
```

As we can see the p-value has increased and the Box cox transformation has helped us to reduce heteroscedasticity to an extent.

Comparing Coefficient of Determination:

Initial Model

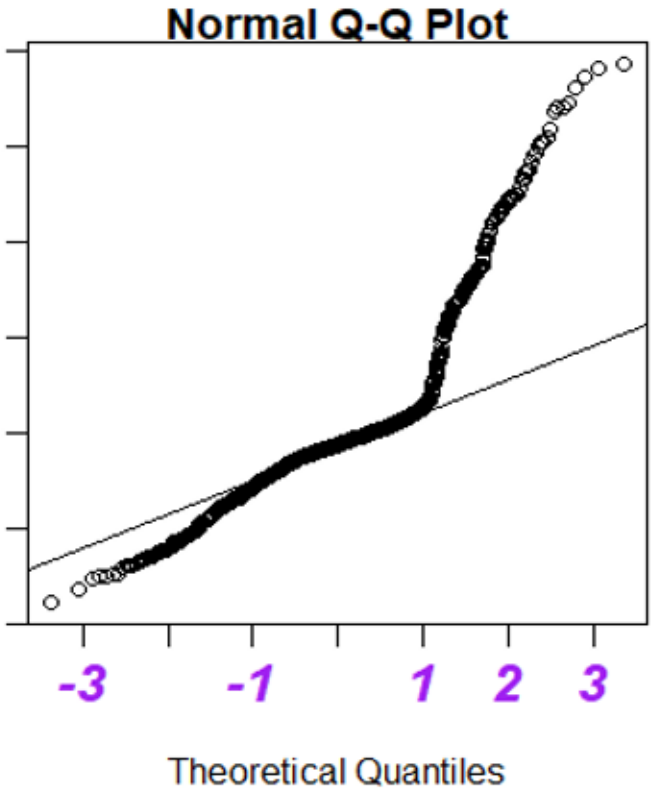
```
Residual standard error: 6062 on 1329 degrees of freedom
Multiple R-squared:  0.7509, Adjusted R-squared:  0.7494
F-statistic: 500.9 on 8 and 1329 DF,  p-value: < 2.2e-16
```

Transformed Model:

```
Residual standard error: 1.589 on 1331 degrees of freedom
Multiple R-squared:  0.7749, Adjusted R-squared:  0.7739
F-statistic: 763.7 on 6 and 1331 DF,  p-value: < 2.2e-16
```

It can clearly be seen that there is a significant improvement in R^2 and AdjR^2 of the transformed model as compared to the original model.

Normal Q-Q plot after Box -Cox Transformation



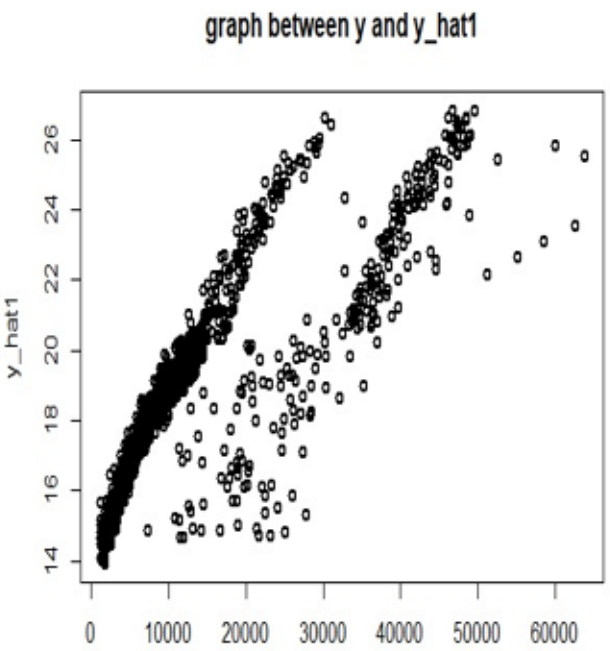
We can clearly see improvement in the data as more observations are now alligned with the straight line.

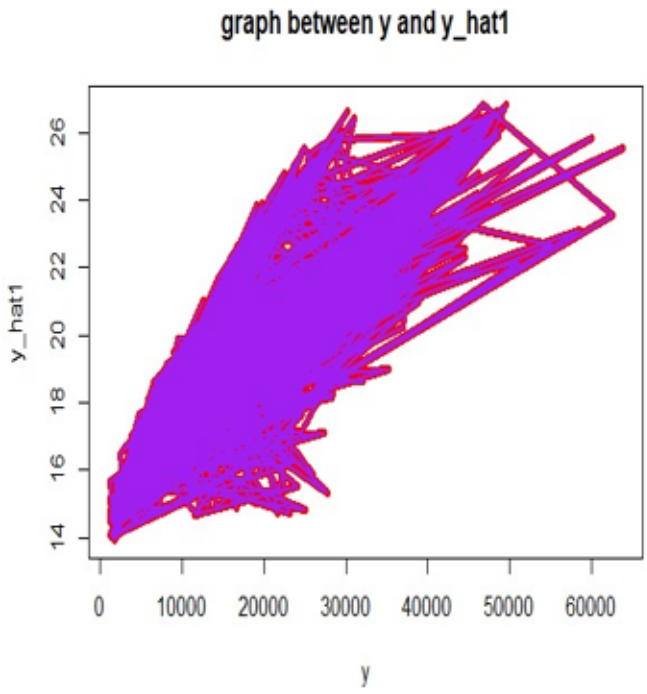
Although after applying the Shapiro- Wilk Normality test on the the transformed model we come to the conclusion that data is still not fully normally distributed but we can assume it as normal for further analysis.

```
        Shapiro-Wilk normality test

data:  residual
W = 0.89911, p-value < 2.2e-16
```

Graph between Fitted and Observed Response(After Transformation):





Conclusion

We could see that above regression model does a pretty good job in predicting the medical expenses for a person using attributes like age,BMI, children,smoker etc.Looking at the coeffiecients directly shows that the most important attributes are the number of children , if the person is a smoker or not and if the person is from North East Region or not.

R Code

```
rm(list=ls())
library(readr)
Insurancefinal_ <- read_csv("C:/Users/hp/Desktop/Insurancefinal .csv")
View(Insurancefinal_)
names(Insurancefinal_)
#assinging values:

y<-Insurancefinal_$Expenses
x1<-Insurancefinal_$Age
x2<-Insurancefinal_$Sex
x3<-Insurancefinal_$BMI
x4<-Insurancefinal_$Children
x5<-Insurancefinal_$Smoker
x6<-Insurancefinal_$SE
x7<-Insurancefinal_$NW
x8<-Insurancefinal_$NE
## Fitting full model

fitt<-lm(y~x1+x2+x3+x4+x5+x6+x7+x8,data=Insurancefinal_)
summary(fitt)
anova(fitt)
## Forming Correlation matrix

x<- cbind(x1,x2,x3,x4,x5,x6,x7,x8)
frame <- data.frame(x1,x2,x3,x4,x5,x6,x7,x8)
```

```
frame
xx=cor(frame)
xx #cor matrix
library(corrplot)
## checking correlation between regressors variables and response variable
corrplot(cor(cbind(y,frame)),method="number")
## Multicollinearity

## installing car package for VIF
#install.packages("car")
library(car)
## checking VIF
vif(fitt)
#since all VIF`s are less than 5`therefore regressors are free from multicollinearity.

## installing mctest package for variance decomposition proportion
#install.packages("mctest")
library(mctest)
eigprop(fitt)
#since maximum condition index is 17.4427 which is > 15 but there are no such
#corresponding subset of regressors which are > 0.5
#hence our model is free from multicollinearity

## Plotting boxplot and density for regressors

par("mar")  #current margin
par(mar=c(1,1,1,1)) #setting margin for plots
par(col.axis="purple", font.axis=4, cex.axis=1.5) # showing ticks
```

```
boxplot(x1,las=0)
hist(x1)
plot(density(x1),las=0)
polygon(density(x1),col= "purple",las=1)

boxplot(x2,las=0)
plot(density(x2))
polygon(density(x2),col= "green",las=0)

boxplot(x3,las=0)
hist(x3)
plot(density(x3))
polygon(density(x1),col= "purple",las=0)

boxplot(x4,las=0)
hist(x4)
plot(density(x4),las=0)
polygon(density(x4),col= "purple",las=0)

boxplot(x5,las=0)
hist(x5)
plot(density(x5),las=0)
polygon(density(x5),col= "purple",las=0)

boxplot(x6,las=0)
hist(x6)
plot(density(x6),las=0)
polygon(density(x6),col= "purple",las=0)
```

```
boxplot(x7,las=0)
hist(x7)
plot(density(x7),las=0)
polygon(density(x7),col= "purple",las=1)

boxplot(x8,las=0)
hist(x8,las=0)
plot(density(x8),las=0)
polygon(density(x8),col= "purple",las=0)

## Checking skewness of regressors
#install.packages("moments")
library(moments)
skewness(x1)
skewness(x2)
skewness(x3)
skewness(x4)
skewness(x5)
skewness(x6)
skewness(x7)
skewness(x8)
##values of skewness fall between ??? 3 and + 3
# of all regressor which is indication of acceptable skewness
#hence we need not to check outliers

## Graphs for checking heteroscedasticity
#install.packages("lmtest")
```

```
library(lmtest)
res <- residuals(fitt)
y_hat <- fitted(fitt)
cor(y_hat,res) # cor()= 0.00000000000000000846319= 8.46319e-17 almost zero (approx.)
        # Hence, our assumption on homoscedastic error holds true.
plot(res,y_hat,las=1)
mean(res)

#res vs predictors graph of residuals against independent variables
plot(x1,res,las=0)
plot(x2,res,las=0)
plot(x3,res)#The plot is close to a outward opening funnel
plot(x4,res,las=0)
plot(x5,res,las=0)
plot(x6,res,las=0)
plot(x7,res,las=0)
plot(x8,res,las=0)
#which suggests the following relationship:
#Var(ei) ??? ??2BMIi
#To ensure our usual error assumptions, we need to account for this and make changes to our model.
#The way we can accomodate such a relationship is by dividing the model on both sides by the BMI
#regressor. So, our new model is

#Breusch-Pagan Test for heteroscedasticity
bptest(fitt)
```

```
fitt<-lm(y~x1+x2+x3+x4+x5+x6+x7+x8)# model before applying weight=1/[BMI(j)]
summary(fitt)
#model after applying weights
fitt11<- lm(y~x1 + x2 + x3 + x4 + x5+x6+x7+x8,weights=1/x3)
summary(fitt11)
bptest(fitt11)
#log transformation in expenses
fitt2<-lm(log(y)~x1+x2+x3+x4+x5+x6+x7+x8)
summary(fitt2)
bptest(fitt2)

#checking normality
qqnorm(res,las=1)
qqline(res,las=1)
shapiro.test(res)

#variable selection
library(MASS)
step(fitt,direction="both")
## Variable Selection Akaike Information criteria corrected

#install.packages("AICcmodavg")
library(AICcmodavg)
AICc(fitt)

## AICc=27115.59

#final model after variable selection
```

```
final_model=lm(formula = y ~ x1 + x3 + x4 + x5 + x7 + x8,data=Insurancefinal_)
summary(final_model)
bptest(final_model)

##Calculation of MSE
mean(fitt$residuals^2)

## prediction of data
cor(y,y_hat)
plot(y,y_hat, xlab="y", ylab="y_hat", main="graph between y and y_hat")
## Graph between observed and fitted values

plot(y,y_hat, type="l", lwd=5, xlab="y", ylab="y_hat", main="graph between y and y_hat")

lines(y, y_hat, col="red", lwd=2)

lines(y, y_hat, type="b", col="purple",pch=19)

#final model after variable selection
#
final_model=lm(formula = y ~ x1 + x3 + x4 + x5 + x7 + x8,data=Insurancefinal_)

library(lmtest)
residual<- residuals(final_model)
y_hattt <- fitted(final_model)
cor(y_hattt,residual)

#transforming data
```

```
library(MASS)
bc=boxcox(final_model)
  (lambda=bc$x[which.max(bc$y)])
#lamda=0.1414141
new_model=lm(((y^lambda-1)/lambda)~x1+x3+x4+x5+x7+x8)

library(lmtest)
residual_<- residuals(new_model)
y_hat_new<- fitted(new_model)
cor(y_hat_new,residual_)
summary(new_model)

bptest(new_model)
shapiro.test(residual)

#define plotting area
op <- par(pty = "s", mfrow = c(1, 2))

#Q-Q plot for original model
qqnorm(final_model$residuals)
qqline(final_model$residuals)

#Q-Q plot for Box-Cox transformed model
qqnorm(new_model$residuals)
qqline(new_model$residuals)

#display both Q-Q plots
```

```
par(op)


y_hat1 <- fitted(new_model)

plot(y,y_hat1,lwd=2, xlab="y", ylab="y_hat1", main="graph between y and y_hat1")

## Graph between observed and fitted values


plot(y,y_hat1, type="l", col="red", lwd=5, xlab="y", ylab="y_hat1", main="graph between y
and y_hat1")


lines(y, y_hat1, col="purple", lwd=2)


lines(y, y_hat, xlab="y", ylab="y_hat1", main="graph between y and y_hat1", lwd=2,
pch=19)


#CONCLUSION: after transformation data:  Residual standard error: 1.589 on 1331 degrees
of freedom

#Multiple R-squared:  0.7749,          Adjusted R-squared:  0.7739

#F-statistic: 763.7 on 6 and 1331 DF,  p-value: < 2.2e-16

#new_model


#BP = 55.176, df = 6, p-value = 4.271e-10    Shapiro-Wilk normality test

                                    #data:  residual

                                   #W = 0.89911, p-value < 2.2e-16




#visualizing the response varible data

hist(y, freq = FALSE)

lines(density(y))
```

```
library("fitdistrplus")
descdist(y, discrete = FALSE)

normal_dist <- fitdist(y, "norm")
plot(normal_dist)
```