Report Summary –

Outcome of the Regression Techniques on Resales of Homes Data

Date: 29th January 2018

Submitted by:

Name	SMS ID	SID
Parthasarathi Chatterjee	118284	TA17021

A. Introduction

A data set (regression_1.csv – Refer Annexure) is provided on the Resales of Homes data. The data set consists of resale home price transactions with 8 variables having the details like Price, Sq Ft., Age, Feats, NE, CUST, COR and Tax.

The requirement is to fit a Multiple Regression Model for finding the determinants of the reselling price of a house given a set of predictors.

B. Approach

Given the data set, a multiple regression model needs to be fitted with resale price of homes as the dependent variable and all other variables as predictors. In addition to the fitting a multiple regression model, the heteroscedasticity, multicollinearity and the violation of normality assumption in the fitted model is also checked and the necessary remedies has been suggested so as to achieve an optimum fitted model

C. Method: Multiple Regression

Step 1: Data Preparation by transforming the values of the Feats variable with dummy variables (binary).

Reason: The numbers provided in the Feats variable are categorical in nature and hence transformed by dummy variable (binary)

Reference R Code:

```
setwd("C:\\Software\\xlri\\Regression\\Assignment")
library(car)
library(lmtest)
library(tseries)
library(sandwich)
library(stats)
library(MASS)
library(faraway)
realestateprice<-read.csv("regression_1.csv",header=TRUE, na.strings=c("*",'NA'))
str(realestateprice)
head(realestateprice)
for(level in unique(realestateprice$Feats)-1){

  realestateprice[paste("Feats",    gsub("-","_", gsub(" ","_",level, fixed=TRUE), fixed=TRUE), sep = "_")] <-
ifelse(realestateprice$Feats == level, 1, 0)

}
realestateprice$Feats__1<-NULL
realestateprice$Feats<-NULL
```

Step 2: **Data Preparation by Imputing missing values for the variables Age and Tax**

Reason: The missing values needed to be imputed for the variables Age and Tax. The following process has been applied for imputation:

(a) The imputation in case of Age variable is done by taking the mean of the available values in the variable.

(b) The imputation in case of Tax is done by Regressing the Tax variable on the other predictor variables and generating a model to achieve the value to be imputed in Tax variable.

Reference R Code:

The R Code for Imputation of Age variable

```
cor(realestateprice,use="complete.obs")
realestateprice$Age[is.na(realestateprice$Age)]=round(mean(realestateprice$Age[!is.na(realestateprice$Age)
],na.rm=FALSE),0)
```

The R Code for Imputation of Tax variable

```
lmtax<-lm(TAX~ SQ.FT+Age+NE+CUST+COR+Feats_7+Feats_6+Feats_5+Feats_3+Feats_4+Feats_2
    +Feats_1+Feats_0 ,data=realestateprice )
lmtax
```

Output of the above model

```
Call:
lm(formula = TAX ~ +SQ.FT + Age + NE + CUST + COR + Feats_7 +
    Feats_6 + Feats_5 + Feats_3 + Feats_4 + Feats_2 + Feats_1 +
    Feats_0, data = realestateprice)

Coefficients:
(Intercept)        SQ.FT          Age           NE         CUST          COR      Feats_7
  -221.6515       0.4503      -6.3950      46.5077      22.0171     -39.9941     681.8870
    Feats_6      Feats_5      Feats_3      Feats_4      Feats_2      Feats_1      Feats_0
   354.7987     375.1991     337.7245     344.0058     294.3831     349.9263      23.4714

> |
```

#Imputing the coefficients from the model to generate the missing values for the TAX variable

```
realestateprice$k<-0

for(i in 1:nrow(realestateprice))
{
 if(is.na(realestateprice$TAX[i])==TRUE)
 {
  realestateprice$k[i]=realestateprice$k[i]+1
 }
}
#
for (i in 1:nrow(realestateprice))
{
 if(realestateprice$k[i]==1)
```

```
 {
realestateprice$TAX[i]= (-221.6515

                        + 0.4503 *realestateprice$SQ.FT[i]
                        -6.3950*realestateprice$Age[i]
                        +46.5077*realestateprice$NE[i]
                        +22.0171*realestateprice$CUST[i]
                        -39.9941*realestateprice$COR[i]
                        +681.8870*realestateprice$Feats_7[i]
                        + 354.7987*realestateprice$Feats_6[i]
                        +375.1991*realestateprice$Feats_5[i]
                        +344.0058 *realestateprice$Feats_4[i]
                        +337.7245*realestateprice$Feats_3[i]
                        +294.3831 *realestateprice$Feats_2[i]
                        +349.9263* realestateprice$Feats_1[i]
                        +23.4714* realestateprice$Feats_0[i]
   )
 }
}
realestateprice$TAX<-round(realestateprice$TAX,0)
realestateprice$k<-NULL
```

Step 3: Writing back the data set to the working directory.

Reason: The data set once transformed with dummy variables for the variable Feats and imputed with values for the variables Age and TAX has been written back to the working directory for further process.

The file Imputed Data_1.csv is attached in the Annexure for reference

Reference R Code:

```
write.csv(realestateprice,"Imputed Data_1.csv")
```

Step 4: Building the multiple regression model on the basis of the imputed data

Reference R Code:

```
lmprice<-lm(Price~SQ.FT+Age+NE+CUST+COR+TAX+Feats_7+Feats_6+Feats_5+Feats_3+Feats_4+Feats_2

       +Feats_1+Feats_0 ,data=realestateprice )


summary(lmprice)
```

Output of the above summary () command

```
Call:
lm(formula = Price ~ SQ.FT + Age + NE + CUST + COR + TAX + Feats_7 +
    Feats_6 + Feats_5 + Feats_3 + Feats_4 + Feats_2 + Feats_1 +
    Feats_0, data = realestateprice)

Residuals:
    Min      1Q  Median      3Q     Max
-539.51  -86.94   -3.26   68.56  545.06
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 212.54857  190.94112   1.113  0.26825
SQ.FT         0.21564    0.07218   2.988  0.00352 **
Age          -0.14455    1.99665  -0.072  0.94243
NE            7.44853   36.81947   0.202  0.84009
CUST        125.74816   47.50376   2.647  0.00941 **
COR         -54.31730   43.12211  -1.260  0.21068
TAX           0.68517    0.13470   5.087 1.66e-06 ***
Feats_7      11.69593  262.37956   0.045  0.96453
Feats_6      -0.15777  191.58367  -0.001  0.99934
Feats_5     -84.75506  189.47235  -0.447  0.65559
Feats_3     -64.40592  185.05882  -0.348  0.72854
Feats_4     -78.15470  185.14314  -0.422  0.67382
Feats_2     -94.95114  185.95695  -0.511  0.61073
Feats_1     -70.08806  194.34225  -0.361  0.71911
Feats_0     -11.95310  218.36154  -0.055  0.95645
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 172.5 on 102 degrees of freedom
Multiple R-squared:  0.8192,   Adjusted R-squared:  0.7944
F-statistic: 33.01 on 14 and 102 DF,  p-value: < 2.2e-16
```

Note :

a) **The transformation of the Feats variable and imputation of the variables Age and TAX and building the regression model is also executed in excel. (Refer : Section D -Conclusion)**
b) **All the predictors variables in the above model seemed to be significant from the business perspective and hence retained in the model. However we will check specific issues like Multicollinearity, Heteroscedasticity and violation of normality in the subsequent steps and address them.**

Step 5: Checking Multicollinearity in the Model

Reason: To check the multicollinearity in the model. i.e. to check if any of the predictor variables are themselves linearly correlated

Reference R Code:

vif(lmprice)

Output of the above vif () command

```
SQ.FT       Age         NE       CUST        COR        TAX    Feats_7    Feats_6
Feats_5
 5.569805  1.440567  1.184372  1.574839  1.116144  6.373824  2.293464
9.191943 12.021651
```

```
  Feats_3   Feats_4   Feats_2   Feats_1   Feats_0 
24.509898 31.324873 16.048708  8.352165  3.149590 
```

Building the Linear Model after removing Multicollinearity

lmpricefit <- lm(Price~.

-Feats_5
-Feats_4
-Feats_3
-Feats_2
,data=realestateprice)

summary(lmpricefit)

```
Call:
lm(formula = Price ~ . - Feats_5 - Feats_4 - Feats_3 - Feats_2, 
    data = realestateprice)

Residuals:
    Min      1Q  Median      3Q     Max 
-546.86  -85.12    0.00   71.66  556.80 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 134.29288   66.27669   2.026  0.04525 *  
SQ.FT         0.22074    0.06939   3.181  0.00193 ** 
Age          -0.12160    1.93825  -0.063  0.95010    
NE            7.48749   35.63280   0.210  0.83397    
CUST        127.54513   43.97647   2.900  0.00453 ** 
COR         -57.15128   41.60907  -1.374  0.17249    
TAX           0.67603    0.12669   5.336 5.43e-07 ***
Feats_6      75.63566   64.49272   1.173  0.24351    
Feats_7      89.28993  180.11490   0.496  0.62111    
Feats_1       6.67923   69.89685   0.096  0.92405    
Feats_0      62.14127  130.47624   0.476  0.63487    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 169.7 on 106 degrees of freedom
Multiple R-squared:  0.8182,	Adjusted R-squared:  0.8011 
F-statistic: 47.72 on 10 and 106 DF,  p-value: < 2.2e-16
```

vif(lmpricefit)

Output of the vif () command on the new model

```
> vif(lmprice)
    SQ.FT       Age        NE      CUST       COR       TAX   Feats_7   Feats_6   Feats_5 
 5.569805  1.440567  1.184372  1.574839  1.116144  6.373824  2.293464  9.191943 12.021651 
  Feats_3   Feats_4   Feats_2   Feats_1   Feats_0 
24.509898 31.324873 16.048708  8.352165  3.149590 
```

Outcome/ Inference:

On running vif, it is found that the vif of the transformed variables like Feats_5,Feats_4,Feats_3, Feats_2 have vif more than 10 and hence dropped from the model.

(Ref : VIF Standard more than 10 implies multicollinearity of that variable)

On dropping the variables, the *Impricefit* is the new model and again vif is checked for the new model. It is observed that there is no further multicollinearity in the new model

Step 6: Checking Heteroscedasticity in the Model by Breusche Pegan Test

Reason : To check the heteroscedasticity in the model , the following hypothesis testing is done :

H0 : The error variances of the model is similar for all observations

H1 : The error variances of the model varies

Reference R Code:

bptest(Impricefit)

Outcome/ Inference:

The below is the result of Breusche Pegan Test. The p-value of the test statistics which is less than 0.05 signifies heteroscedasticity in the model

Output of the Breusche Pegan Test

```
studentized Breusch-Pagan test

data:  lmpricefit
BP = 43.119, df = 10, p-value = 4.735e-06
```

Step 7: **Checking Violation of Normality assumption in the Model by Shapiro Wilk Test**

Reason: To check the violation of Normality assumption in the model, the following hypothesis testing is done

H0: The sample of observations taken from a normally distributed population

H1: The sample of observations has not come from a normally distributed population

Reference R Code:

shapiro.test(resid(lmpricefit))

Outcome/ Inference:

The below is the result of Shapiro Wilk Test. The p-value which is significantly less than 0.05 signifies violation of normality assumption in the model

Output of the Shapiro Wilk Test

```
Shapiro-Wilk normality test

data:  resid(lmpricefit)
W = 0.95037, p-value = 0.0002801
```

Step 8: Remedy/Rectification of Heteroscedasticity and Violation of Normality assumption by Box Cox Transformation and building the multiple regression model once again

Reason: Since both Heteroscedasticity and Violation of Normality Assumption are observed in the model, Box Cox transformation is applied as a remedy for both

Reference R Code:

lamdabx<-boxcox(lmpricefit)

trans_df = as.data.frame(lamdabx)

optimal_lambda = round(trans_df[which.max(lamdabx$y),1],4)

optimal_lambda

Solution Proposed: The Box-Cox transformation is defined as:

T(Y)= $(Y^{\lambda}$-1)/ λ

where Y is the response variable and λ (lambda)is the transformation parameter.

Box-Cox transformation defined above is helpful to define a measure of the normality of the resulting transformation. One measure is to compute the correlation coefficient of a normal probability plot. The correlation is computed between the vertical and horizontal axis variables of the probability plot and is a convenient measure of the linearity of the probability plot (the more linear the probability plot, the better a normal distribution fits the data).

The Box-Cox normality plot is a plot of these correlation coefficients for various values of the λ parameter. The value of λ (lambda) corresponding to the maximum correlation on the plot is then the optimal choice for λ

The optimal λ (lambda) computed as per the above code = -0.101

The λ (lambda) thus calculated is incorporated on the dependent variable (price) to achieve a new model which is

lmprice_model_cox <- lm((((Price ^ optimal_lambda) - 1) / optimal_lambda) ~ SQ.FT +Ag +NE+CUST +COR+TAX
+Feats_7+Feats_6 +Feats_1 +Feats_0 ,data = realestateprice)

summary(lmprice_model_cox)

Output of the above summary () command for the new model that incorporates the optimal lambda

```
Call:
lm(formula = (((Price^optimal_lambda) - 1)/optimal_lambda) ~ 
    SQ.FT + Age + NE + CUST + COR + TAX + Feats_7 + Feats_6 + 
        Feats_1 + Feats_0, data = realestateprice)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.221990 -0.039522  0.004661  0.045042  0.167806 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  4.582e+00  2.811e-02 163.040  < 2e-16 ***
```

```
SQ.FT        1.100e-04  2.943e-05   3.737 0.000302 ***
Age         -1.272e-04  8.219e-04  -0.155 0.877354    
NE           4.949e-03  1.511e-02   0.327 0.743940    
CUST         3.743e-02  1.865e-02   2.007 0.047257 *  
COR         -1.876e-02  1.764e-02  -1.063 0.289991    
TAX          2.587e-04  5.372e-05   4.816  4.9e-06 ***
Feats_7     -2.093e-02  7.638e-02  -0.274 0.784566    
Feats_6      1.891e-03  2.735e-02   0.069 0.944999    
Feats_1     -5.911e-03  2.964e-02  -0.199 0.842303    
Feats_0     -5.579e-02  5.533e-02  -1.008 0.315636    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.07195 on 106 degrees of freedom
Multiple R-squared:  0.8094,	Adjusted R-squared:  0.7914 
F-statistic: 45.01 on 10 and 106 DF,  p-value: < 2.2e-16
```

The above model is again tested for Breusche Pegan Test and Shapiro Wilk test respectively. The results are as follows:

bptest(lmprice_model_cox)

```
	studentized Breusch-Pagan test

data:  lmprice_model_cox
BP = 33.742, df = 10, p-value = 0.0002042
```

shapiro.test(resid(lmprice_model_cox))

```
	Shapiro-Wilk normality test

data:  resid(lmprice_model_cox)
W = 0.98458, p-value = 0.2019
```

Both the tests after Box Cox transformation reflect substantial improvement of p value. Breusche Pegan test after Box Cox transformation shows significant reduction in heteroscedasticity in the model.(p- value before 4.735e-06, **p-value after** 0.0002042)

Shapiro Wilk test after Box Cox transformation shows that the violation of normality assumption is rectified in the model as the p-value after transformation is 0.2019, **which is higher than 0.05**

The plots generated as per the Box Cox transformation are as follows:

plot(lmprice_model_cox)

Plot 1: Displays the normal distribution post box cox transformation deriving lambda

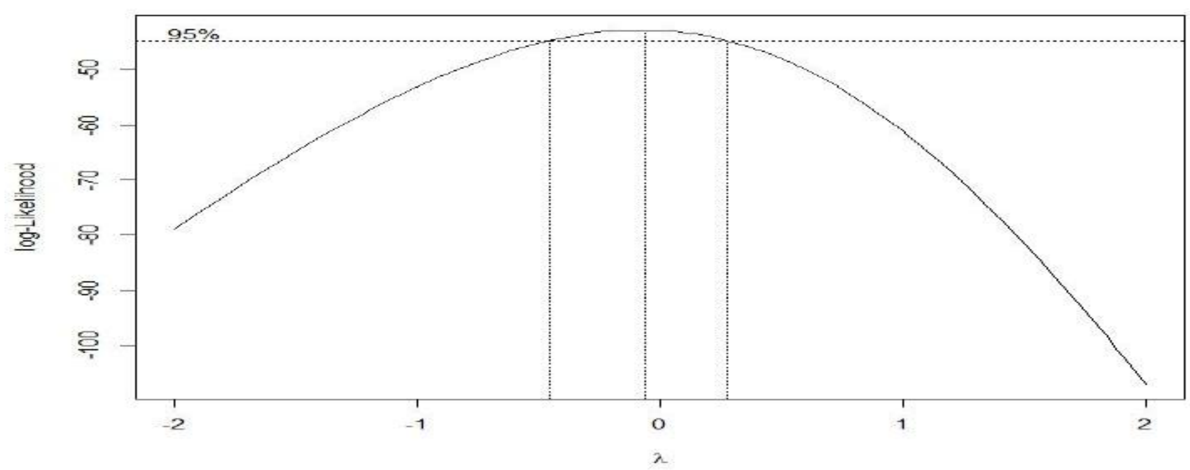

Figure 1: Normal Distribution plot deriving lambda

Plot 2: Displays the relatively uniform distribution of residuals post box cox transformation implying reduction in heteroscedasticity

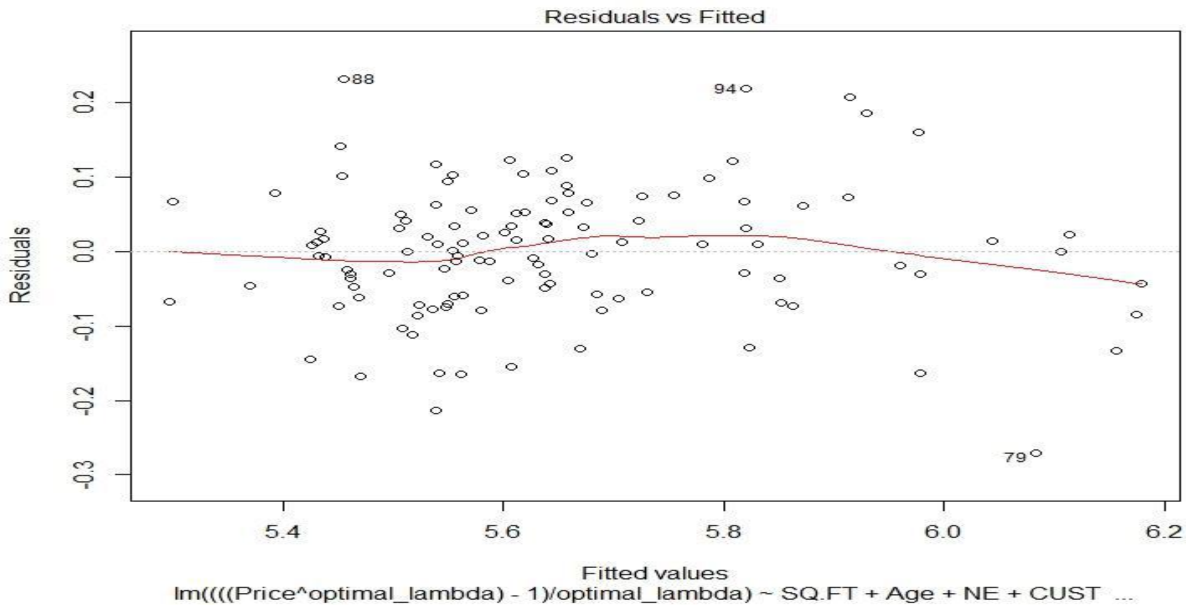

Figure 2: Residual VS. Fitted Values post Box Cox transformation

Plot 3: Displays the improvement in the model as far as the heteroscedasticity is concerned as there is little discernible pattern in the plot and inclination to normal distribution

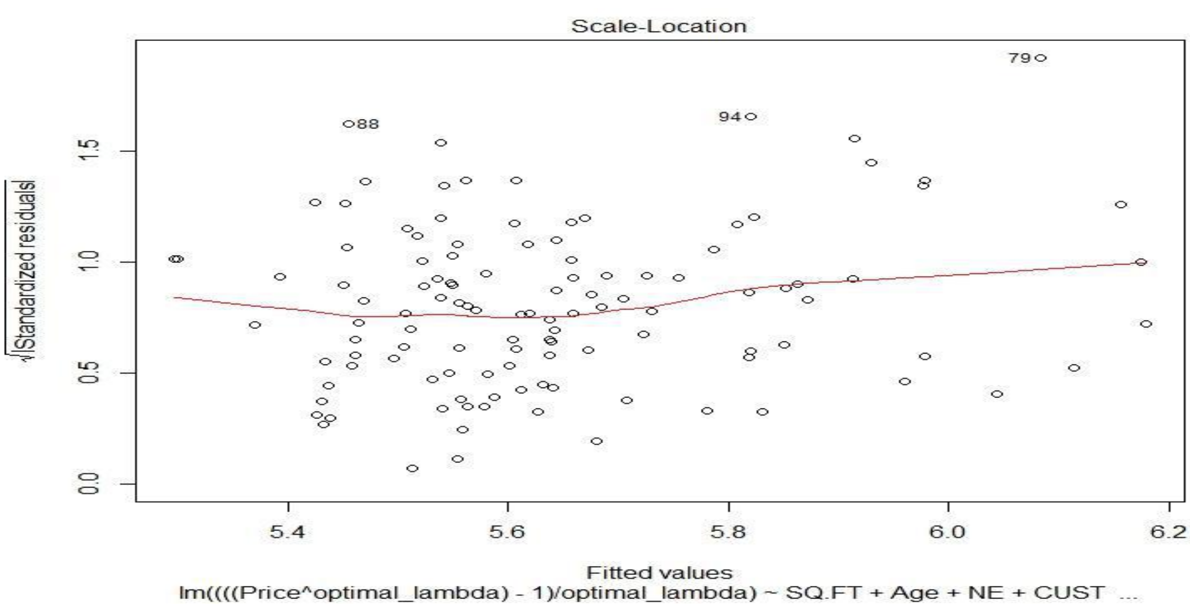

Figure 3: Standardized residual VS Fitted Values

Plot 4: Displays very few observations lying in the range of Cook’s distance implying very few outliers in the model.

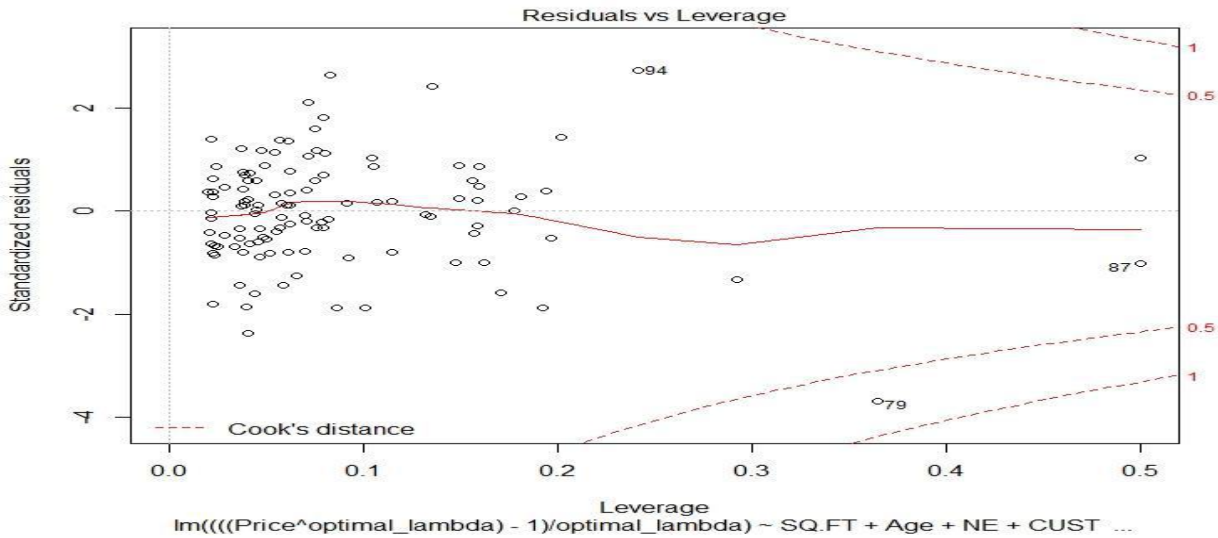

Figure 4: Leverage VS. Standardized residuals post Box Cox transformation

D. Conclusion

We conclude that SQ.FT, Tax and CUST are significant predictors of the Price of the house.

The final equation as per the model is as follows(using R)

Price^(optimal_lambda) - 1/optimal_lambda

=4.582
+0.00011*SQ.FT
-0.0001272*Age
+ 0.004949*NE
+0.03743*CUST
-0.01876*COR
+0.0002587*TAX
-0.02093*Feats_7
+0.001891*Feats_6
-0.005911* Feats_1
-0.05579* Feats_0

The final computation and model (using excel)

Worksheet in C
Software xlri Regres

ANNEXURE

1. **Working Data Set**

regression_1.csv

2. **Imputed Data Set**

Imputed Data_1.csv

3. **R code**

R_Code.txt