IMSE 514 – MULTIVARIATE STATISTICS
HOMEWORK 2

SURESH OOTY

Data Description: HW2-data.txt

The insurance company due to various reasons may sometime decline the house insurance renewal applications. Some researchers from a nonprofit organization collected house insurance data from different cities and tried to investigate the potential factors of declination of house renewal applications. The data show some important statistics that describe the area where the house insurance applicants are located.

- Flood: During the raining season, 1: unlikely to have flood; 2: occasionally have flood; 3: very likely to have flood
- MinorityPop (%): Percentage of minority population
- FireReport (%): Average fire incident per 100 units of house building
- CrimeRate (%): Average crime report per 1000 population
- HouseAge: Average house building age
- Income ($k): Median household income.
- Declination (%): Percentage of declinations in the investigating area

Please conduct analysis as thoroughly as you can base on what we have discussed in the class so far. Make your report as professional as possible. Think about what information you would like to include in the report. Don’t forget to discuss your analytical results.

Solution:

Data preparation & Linear Regression Model

```
> hw2data<-read.csv("./HW2-data.txt",header=T,sep="\t")
> view(hw2data)
> hw2data
   Flood MinorityPop FireReport CrimeRate HouseAge Income Declination
1      2        54.0       34.1      68.0     52.6  82.31        0.30
2      2         4.9       11.0      75.0     42.6 214.80        0.02
3      1         7.1        6.9      18.0     78.5 111.04        0.02
4      3         5.3        7.3      31.0     90.1 106.94        0.40
5      1        21.5       15.1      25.0     89.8  96.31        1.10
6      1        43.1       29.1      34.0     82.7  79.95        1.90
7      1         1.1        2.2      14.0     40.2 137.22        0.02
8      3         1.0        5.7      11.0     27.9 162.50        0.02
9      2         1.7        2.0      11.0      7.7 136.86        0.02
10     1         1.6        2.5      22.0     63.8 124.05        0.02
11     2         1.5        3.0      17.0     51.2 121.98        0.02
12     3         1.8        5.4      27.0     85.1 116.00        0.02
13     3         1.0        2.2       9.0     44.4 127.65        0.02
14     3         2.5        7.2      29.0     84.2 110.84        0.20
15     2        13.4       15.1      30.0     89.8 105.10        0.80
16     1        59.8       16.5      40.0     72.7  97.84        0.80
17     1        94.4       18.4      32.0     72.9  73.42        1.80
18     3        86.2       36.2      41.0     63.1  65.65        1.80
19     1        50.2       39.7     147.0     83.0  74.59        0.90
20     3        74.2       18.5      22.0     78.3  80.14        1.90
21     1        55.5       23.3      29.0     79.0  81.77        1.50
22     2        62.3       12.2      46.0     48.0  82.12        0.60
23     2        10.0        6.2      29.0     60.4 117.44        0.02
24     3        22.2        9.5      44.0     76.5  93.23        0.10
25     2        19.6       10.5      36.0     73.5  99.48        1.20
26     3        17.3        7.7      37.0     66.9 106.56        0.50
27     2        24.5        8.6      53.0     81.4  97.30        0.70
28     3         4.4        5.6      23.0     71.5 112.30        0.30
29     1        46.2       21.8       4.0     73.1  83.30        1.30
30     1        99.7       21.6      31.0     65.0  55.83        0.90
31     1        73.5        9.0      39.0     75.4  85.64        0.40
32     3        10.7        3.6      15.0     20.8 121.02        0.02
33     2         1.5        5.0      32.0     61.8 118.76        0.02
34     1        48.8       28.6      27.0     78.1  97.42        1.40
35     1        98.9       17.4      32.0     68.6  75.20        2.20
36     1        90.6       11.3      34.0     73.4  73.88        0.80
37     2         1.4        3.4      17.0      2.0 238.42        0.02
38     2        71.2       11.9      46.0     57.0 110.40        0.90
39     2        94.1       10.5      42.0     55.9 103.32        0.90
40     3        66.1       10.7      43.0     67.5 109.08        0.40
41     1        36.4       10.8      34.0     58.0 111.56        0.90
42     2         1.0        4.8      19.0     15.2 133.23        0.02
43     3        42.5       10.4      25.0     40.8 129.60        0.50
44     3        35.1       15.6      28.0     57.8 112.60        1.00
45     3        47.4        7.0       3.0     11.4 100.80        0.20
46     2        34.0        7.1      23.0     49.2 114.28        0.30
47     1         3.1        4.9      27.0     46.6 137.31        0.02
48     2        23.7        1.5      18.0     22.0 270.20        0.07
49     3        48.2        3.6      29.3     62.6  85.20        0.75
50     1        18.0        7.3      31.2     18.2  73.50        1.20
```

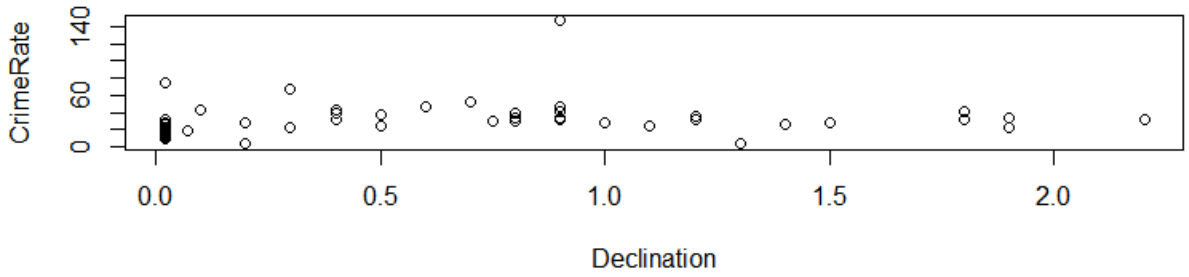
As the Flood variable is not continuous, it was considered as a factor.
Further a step command was used to build the regression model for the Declination.

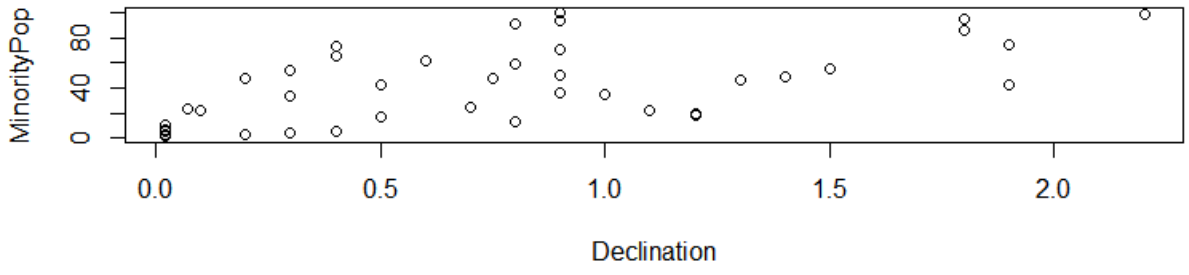
```
> Flood<-as.factor(Flood)
> model1<-lm(Declination~1,data=hw2data)
> model2<-lm(Declination~.,data=hw2data)
> step(model2,data=hw2data,direction="backward")
Start:  AIC=-93.48
Declination ~ Flood + MinorityPop + FireReport + CrimeRate +
    HouseAge + Income

               Df Sum of Sq     RSS     AIC
```

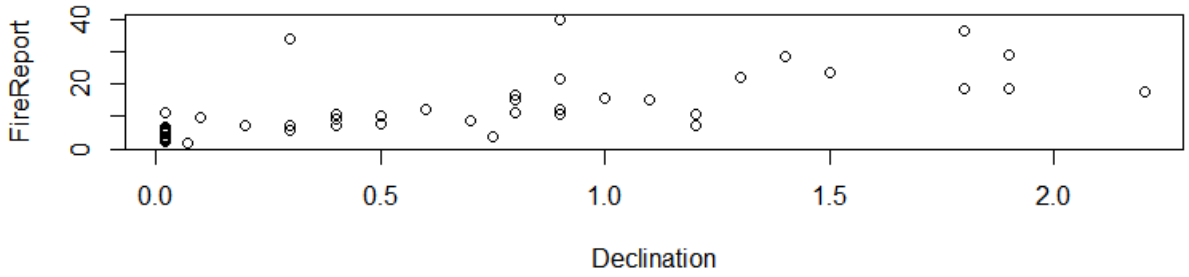
```
- Income       1   0.03888 5.8658 -95.144
- Flood        1   0.12392 5.9509 -94.424
<none>                     5.8269 -93.476
- HouseAge     1   0.30908 6.1360 -92.892
- CrimeRate    1   1.10767 6.9346 -86.775
- MinorityPop  1   1.53054 7.3575 -83.815
- FireReport   1   1.97813 7.8051 -80.862

Step:  AIC=-95.14
Declination ~ Flood + MinorityPop + FireReport + CrimeRate +
    HouseAge

              Df Sum of Sq    RSS     AIC
- Flood        1   0.12178 5.9876 -96.117
<none>                     5.8658 -95.144
- HouseAge     1   0.54137 6.4072 -92.730
- CrimeRate    1   1.20799 7.0738 -87.781
- MinorityPop  1   1.98206 7.8479 -82.589
- FireReport   1   2.18130 8.0471 -81.335

Step:  AIC=-96.12
Declination ~ MinorityPop + FireReport + CrimeRate + HouseAge

              Df Sum of Sq    RSS     AIC
<none>                     5.9876 -96.117
- HouseAge     1    0.5642 6.5518 -93.614
- CrimeRate    1    1.2046 7.1922 -88.951
- MinorityPop  1    2.1018 8.0894 -83.073
- FireReport   1    2.3648 8.3524 -81.474

Call:
lm(formula = Declination ~ MinorityPop + FireReport + CrimeRate +
    HouseAge, data = hw2data)

Coefficients:
(Intercept)  MinorityPop   FireReport    CrimeRate     HouseAge
  -0.092336     0.008053     0.035380    -0.008805     0.005139

> summary(res)

Call:
lm(formula = Declination ~ MinorityPop + FireReport + CrimeRate +
    HouseAge, data = hw2data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.92054 -0.15588 -0.05601  0.15323  1.07030

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.092336   0.148133  -0.623 0.536214
MinorityPop  0.008053   0.002026   3.974 0.000253 ***
FireReport   0.035380   0.008392   4.216 0.000118 ***
CrimeRate   -0.008805   0.002926  -3.009 0.004284 **
HouseAge     0.005139   0.002496   2.059 0.045290 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

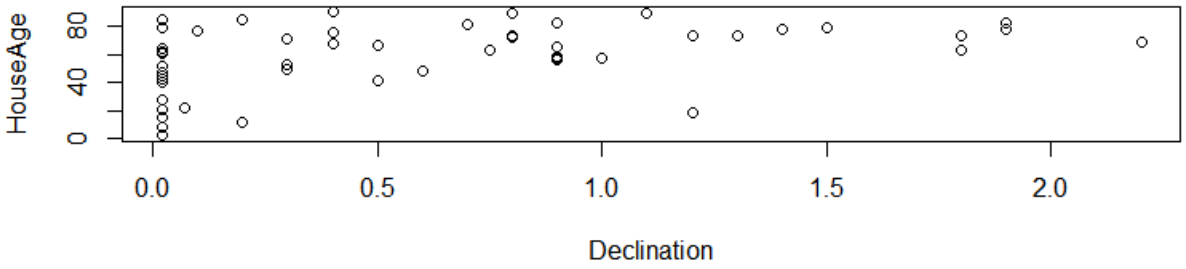
```
Residual standard error: 0.3648 on 45 degrees of freedom
Multiple R-squared:  0.681,    Adjusted R-squared:  0.6526
F-statistic: 24.02 on 4 and 45 DF,  p-value: 1.117e-10
```

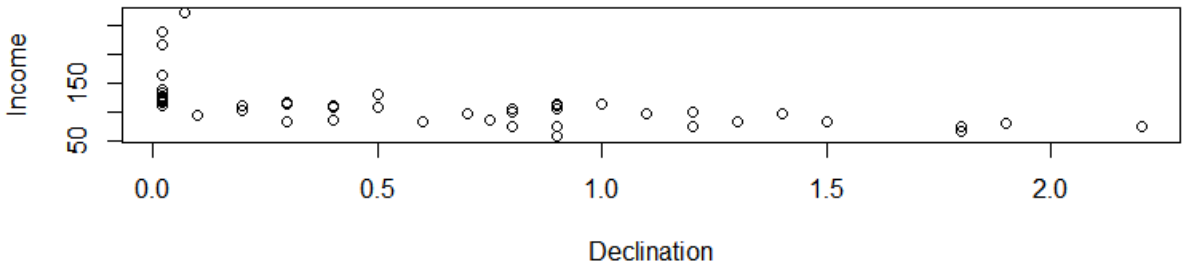
The regression model suggested that variables Minority Population, Fire Report, Crime Rate and House Age significantly influences the Declination of house renewals. However, when the individual variables were plotted against the Response variable, the following observation was made.



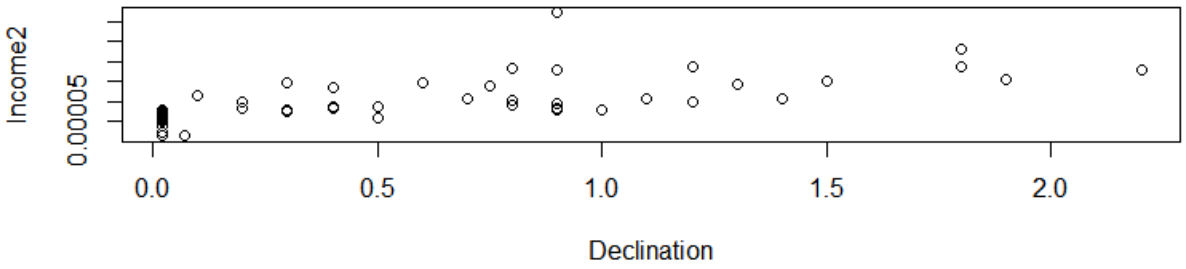








The variable “Income” does not seem to be in linear relationship with Declination. Hence, a transformation approach was adopted using Income = 1 / Income^2 , to make the data look like the following.



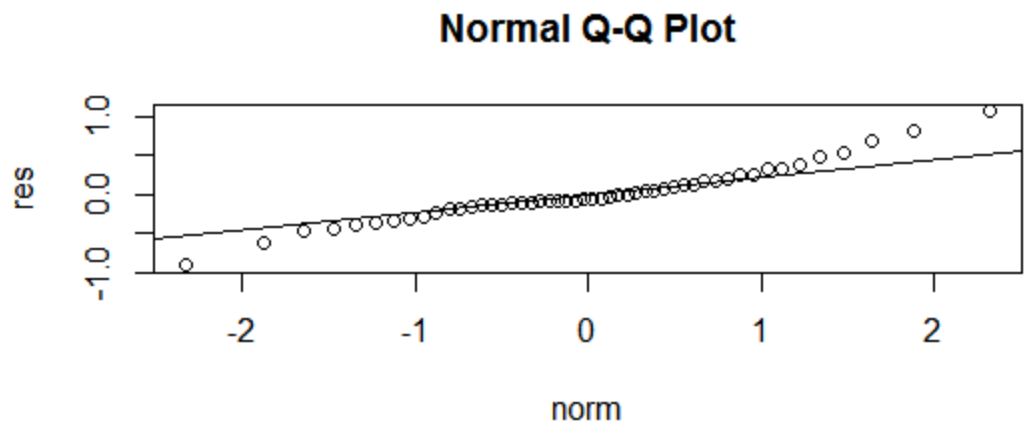
With the transformed data, the regression model was reconstructed to check if there were any influence caused by Income on Declination of house renewal.

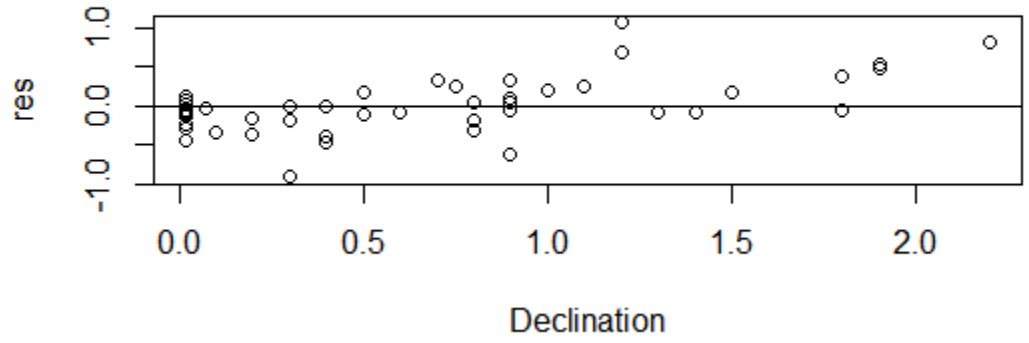
```
> Income2<-1/Income^2
> hw2data$Incnew<-Income2
> model3<-lm(Declination~Income2+HouseAge+CrimeRate+FireReport+MinorityPop+Fl
ood,data=hw2data)
> res1<-step(model3,data=hw2data,direction="backward")
```

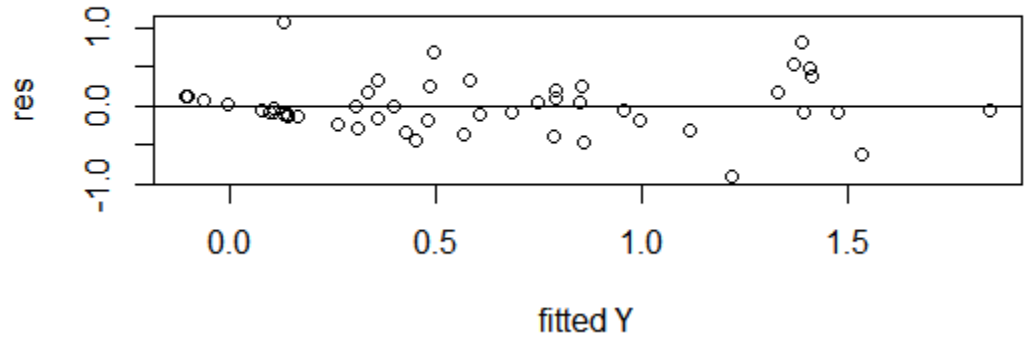
```
Step:  AIC=-96.12
Declination ~ HouseAge + CrimeRate + FireReport + MinorityPop

              Df Sum of Sq    RSS     AIC
<none>                      5.9876 -96.117
- HouseAge     1    0.5642 6.5518 -93.614
- CrimeRate    1    1.2046 7.1922 -88.951
- MinorityPop  1    2.1018 8.0894 -83.073
- FireReport   1    2.3648 8.3524 -81.474
> summary(res1)

Call:
lm(formula = Declination ~ HouseAge + CrimeRate + FireReport +
    MinorityPop, data = hw2data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.92054 -0.15588 -0.05601  0.15323  1.07030

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.092336   0.148133  -0.623 0.536214
HouseAge     0.005139   0.002496   2.059 0.045290 *
CrimeRate   -0.008805   0.002926  -3.009 0.004284 **
FireReport   0.035380   0.008392   4.216 0.000118 ***
MinorityPop  0.008053   0.002026   3.974 0.000253 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3648 on 45 degrees of freedom
Multiple R-squared:  0.681,    Adjusted R-squared:  0.6526
F-statistic: 24.02 on 4 and 45 DF,  p-value: 1.117e-10
```

But this did not change the regression model or the R-squared value. Hence proving that the earlier regression model is valid.
Further checks were made on the residuals.

```
> residuals<-resid(res)
> qqnorm(residuals,ylab="res",xlab="norm")
> qqline(residuals)
> plot(Declination,residuals,ylab="res",xlab="Declination")
> abline(0,0)
> fittedY<-fitted.values(res)
> plot(fittedY,residuals,ylab = "res",xlab="fitted Y")
> abline(0,0)
```



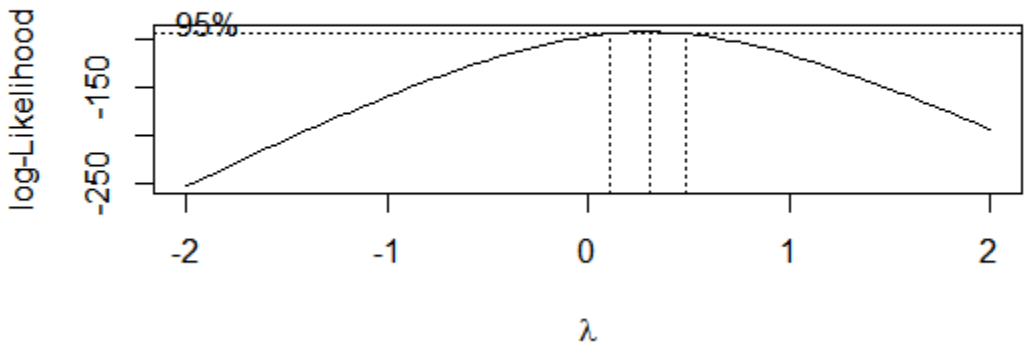


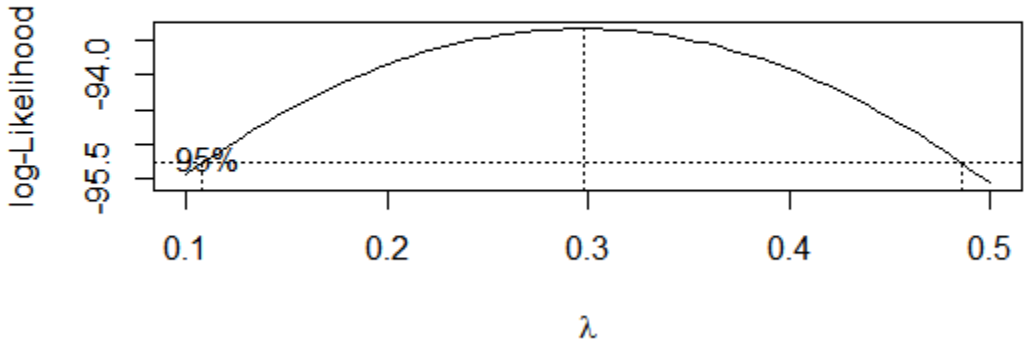


A nonlinear and non-constant variance was noted.

Transformation of response.

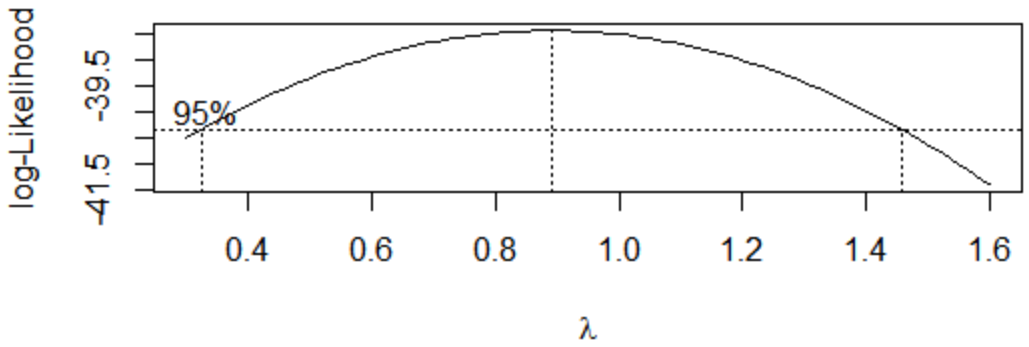
```
> library(MASS)
> bcex1<-lm(Declination~ MinorityPop + FireReport + CrimeRate + HouseAge, dat
a = hw2data)
> boxcox(bcex1,plotit=T)
> boxcox(bcex1,plotit = T,lambda = seq[0.1,0.6,by=0.05])
> boxcox(bcex1,plotit = T,lambda = seq(0.1,0.5,by=0.025))
> boxcox(bcex1,plotit = T,lambda = seq(0.1,0.6,by=0.05))
```



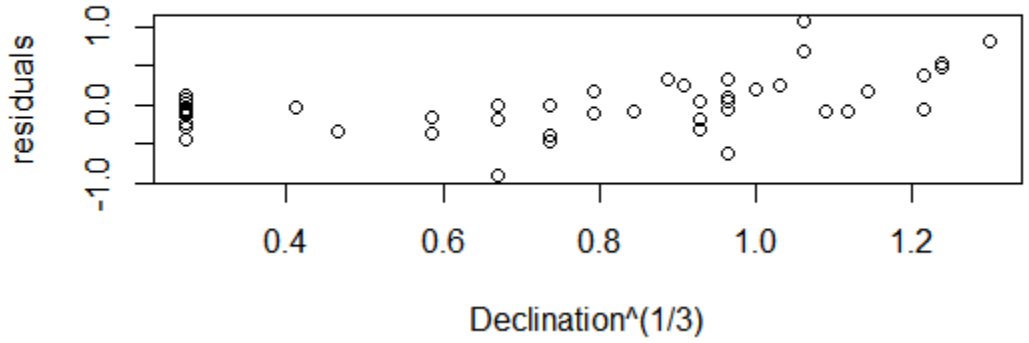


A lambda of 0.3 is noted. Hence the response Declination is transformation to Declination^(1/3) to get the following improved fit.

```
> bcex1new<-lm(Declination^(1/3)~ MinorityPop + FireReport + CrimeRate + Hous
eAge, data = hw2data)
> boxcox(bcex1new,plotit=T)
> boxcox(bcex1new,plotit = T,lambda = seq(0.3,1.6,by=0.1))
```



And the transformed response was validated against the residuals to note independency, hence addressing the concerns over assumptions.



The final Equation:

```
Declination = -0.092336 + HouseAge(0.005139) + CrimeRate (-0.008805) + FireRe
port(0.035380)+ MinorityPop(0.008053)
```