Practical_5_regression_2022

Soumarya Basak

14/05/2022

Transformation Of Variable

Problem 1

Import the data

```
library(readr)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
## v ggplot2 3.3.5     v dplyr   1.0.7
## v tibble  3.1.6     v stringr 1.4.0
## v tidyr   1.1.4     v forcats 0.5.1
## v purrr   0.3.4
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

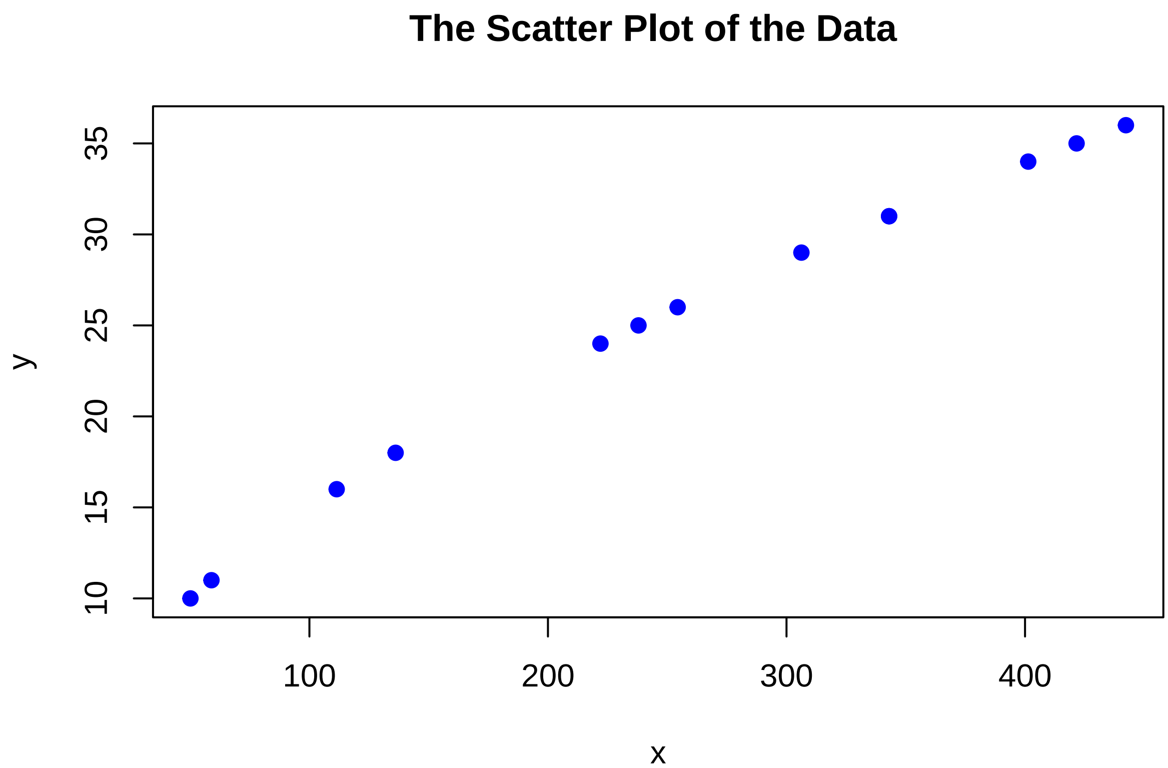
```
library(readxl)
```

```
df <- read.csv("C:\\Users\\souma\\Dropbox\\Mstat_CU\\Sem 2\\Regression_analysis_1\\Data Sets\\problem_s
colnames(df) <- c("y","x")
df
```

```
##     y        x
## 1  10  50.1187
## 2  11  58.9342
## 3  16 111.4305
## 4  18 136.1330
## 5  25 237.9567
## 6  24 222.0031
## 7  26 254.3634
## 8  29 306.2504
## 9  31 343.0164
## 10 31 343.0164
## 11 34 401.3416
## 12 35 421.6146
## 13 36 442.2973
```

Plot of the Data

```
plot(df$x,df$y,pch=19,col='blue',
     xlab="x",ylab = "y",
     main="The Scatter Plot of the Data ")
```



Comment:

Fit a model

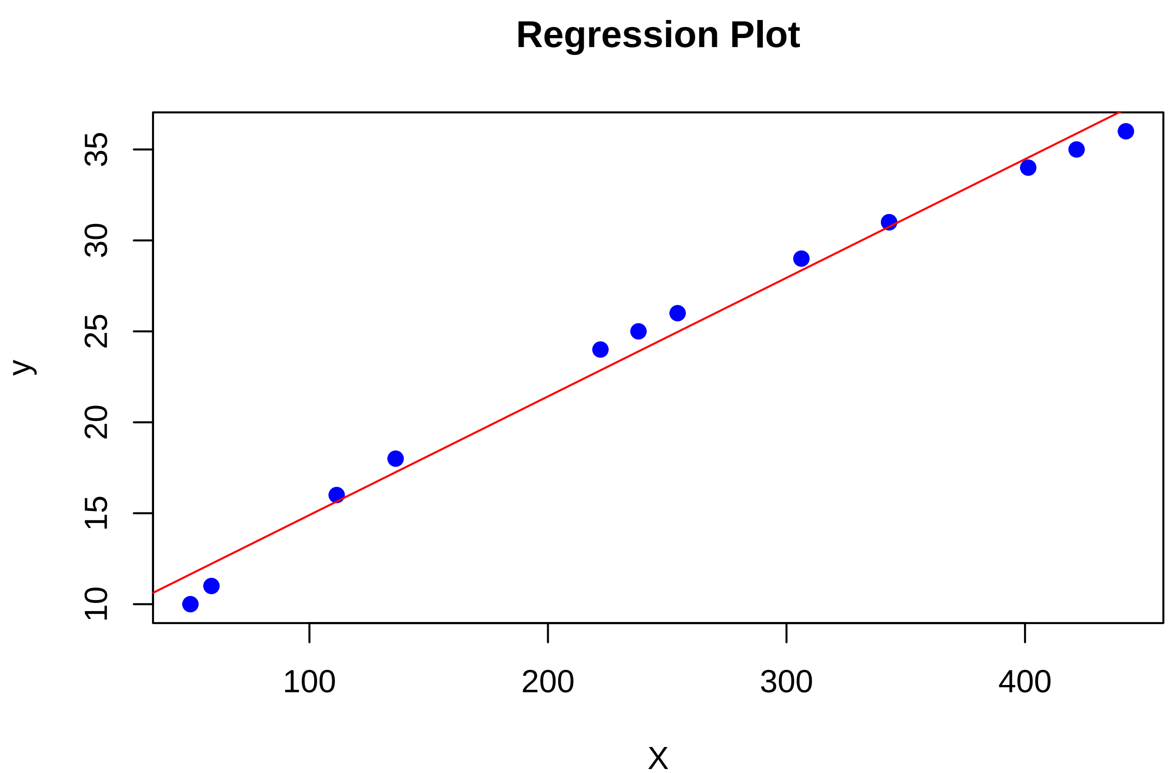
```
model1<- lm(y~x,data = df)
summary(model1)
```

```
##
## Call:
## lm(formula = y ~ x, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.6455 -0.8771  0.2497  0.7440  1.1430
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 8.376417   0.629721   13.30 4.01e-08 ***
## x           0.065227   0.002194   29.73 7.35e-12 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.026 on 11 degrees of freedom
## Multiple R-squared:  0.9877, Adjusted R-squared:  0.9866
## F-statistic:   884 on 1 and 11 DF,  p-value: 7.35e-12
```

Comment: From the above model fit we see that both the coefficients are significant, and also the adjusted R squared is 0.98 and the residual standard error is also low, so it is a quite good fit.

```r
plot(df$x,df$y,pch=19,col='blue',
     xlab = "X",ylab = "y",
     main = "Regression Plot")
abline(model1, col='red')
```



From the plot we also see the plot fits well. So **I am quite satisfied**.

Although, it is a good fit, lets try a Box_Tidwell transformation,

$$x \to x^* = x^{(1/2)}$$

Box-Tidwell transformation

This transformation makes a non-linear model to linear, here the plot is a arc of a parabolic curve, i.e. y^2 depends on x linearly, so we make a square root transformation.

$$x \to x^* = x^{(1/2)}$$

```r
df$x1 <- df$x^(1/2)
df
```

```
##     y        x        x1
## 1  10  50.1187  7.079456
## 2  11  58.9342  7.676861
## 3  16 111.4305 10.556065
## 4  18 136.1330 11.667605
## 5  25 237.9567 15.425845
## 6  24 222.0031 14.899768
## 7  26 254.3634 15.948774
## 8  29 306.2504 17.500011
## 9  31 343.0164 18.520702
## 10 31 343.0164 18.520702
## 11 34 401.3416 20.033512
## 12 35 421.6146 20.533256
## 13 36 442.2973 21.030865
```

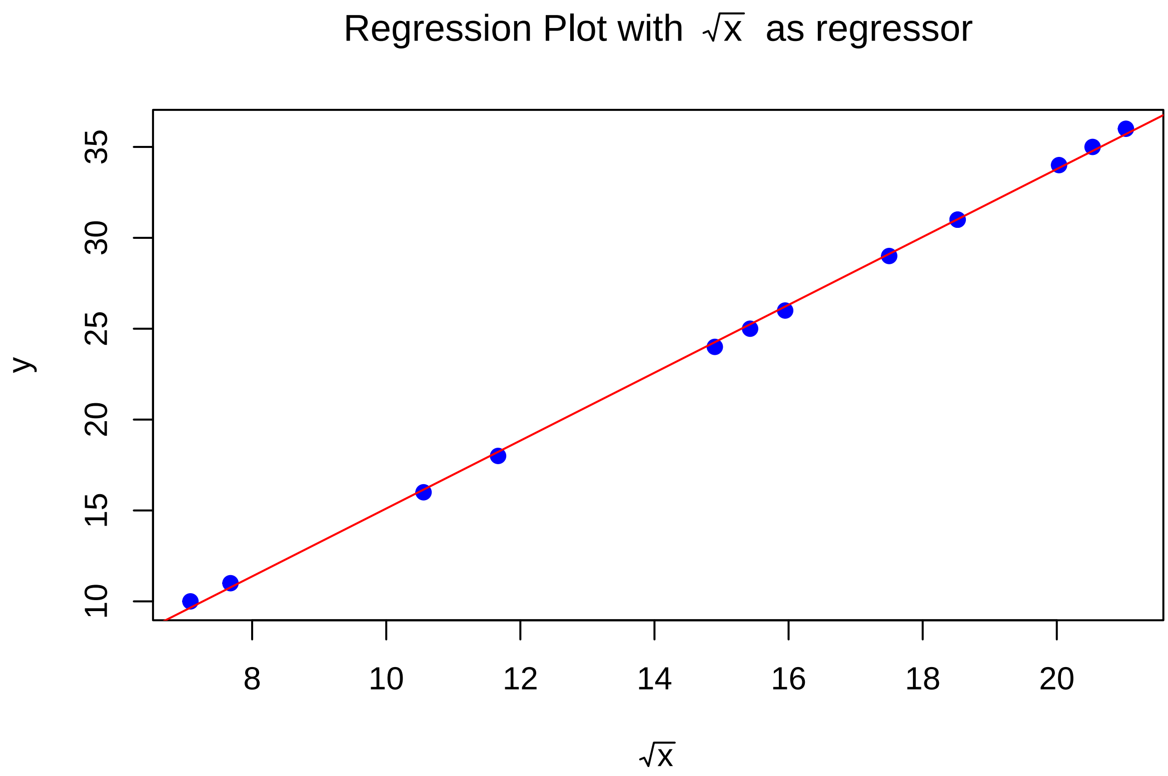
Now regressing y with respect to $x^{1/2}$,

```r
model2<- lm(y~x1,data = df)
summary(model2)
```

```
##
## Call:
## lm(formula = y ~ x1, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.25859 -0.21765 -0.02082  0.22066  0.34613
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.56727    0.22497  -15.86 6.34e-09 ***
## x1           1.86754    0.01406  132.83  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2311 on 11 degrees of freedom
## Multiple R-squared:  0.9994, Adjusted R-squared:  0.9993
## F-statistic: 1.764e+04 on 1 and 11 DF,  p-value: < 2.2e-16
```

For this model the coefficient are also significant, and the Adjusted R-squared value is 0.99, which is larger than previous value and the residual standard error is also lower than the previous one. SO, the transformation more stabilizes the model.

```r
plot(df$x1,df$y,pch=19,col="blue",
     xlab = expression(sqrt(x)),ylab = "y",main = expression( "Regression Plot with "~ sqrt(x) ~" as reg
abline(model2, col='red')
```



Problem 2

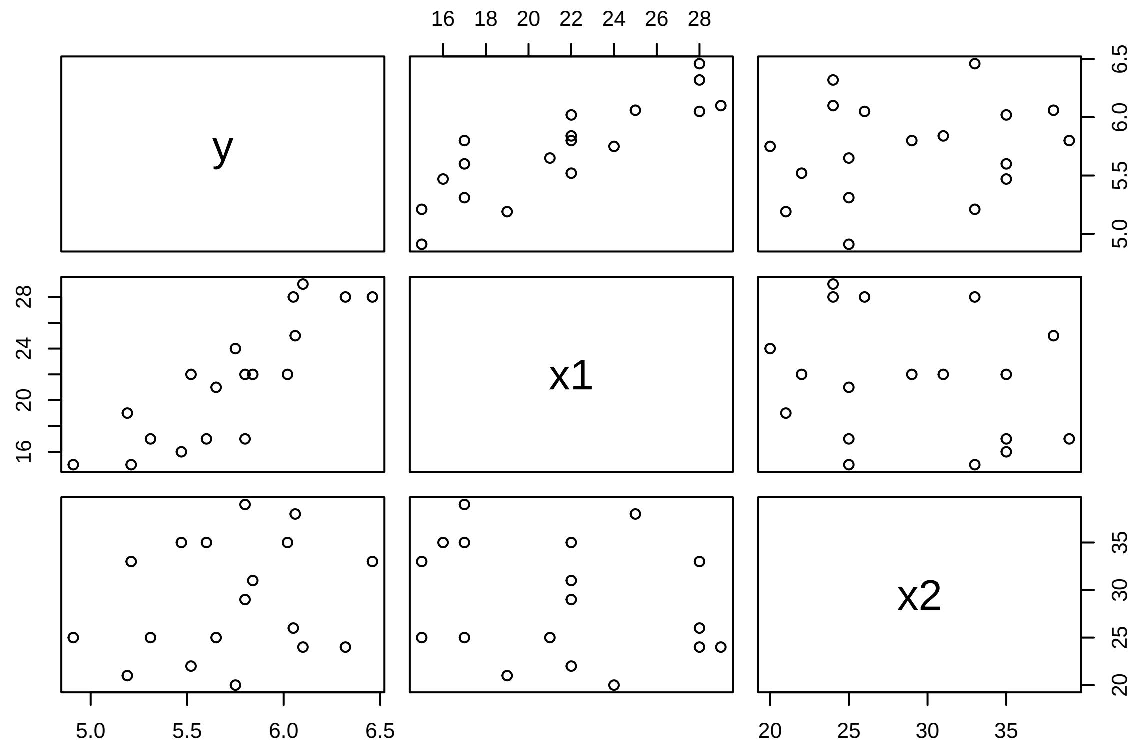
import the data

```
df2<- read.csv("C:\\Users\\souma\\Dropbox\\Mstat_CU\\Sem 2\\Regression_analysis_1\\Data Sets\\problem_s
colnames(df2)<- c("y","x1","x2")
df2
```

```
##       y x1 x2
## 1  6.46 28 33
## 2  5.65 21 25
## 3  6.02 22 35
## 4  5.60 17 35
## 5  5.47 16 35
## 6  6.32 28 24
## 7  5.80 22 29
## 8  5.52 22 22
## 9  6.05 28 26
## 10 6.10 29 24
## 11 5.31 17 25
## 12 5.21 15 33
## 13 5.19 19 21
## 14 5.80 17 39
## 15 5.84 22 31
## 16 6.06 25 38
```

```
## 17 5.75 24 20
## 18 4.91 15 25
```

```r
# variables plot
plot(df2)
```



Fit the linear Regression

```r
model3<- lm(y~ x1+x2, data = df2)
summary(model3)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2, data = df2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.23721 -0.09533  0.02282  0.07600  0.20155
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3.044288   0.223767  13.605 7.64e-10 ***
## x1          0.083508   0.006377  13.096 1.30e-09 ***
## x2          0.030664   0.004999   6.135 1.91e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 0.1213 on 15 degrees of freedom
## Multiple R-squared:  0.9234, Adjusted R-squared:  0.9132
## F-statistic:  90.4 on 2 and 15 DF,  p-value: 4.288e-09
```

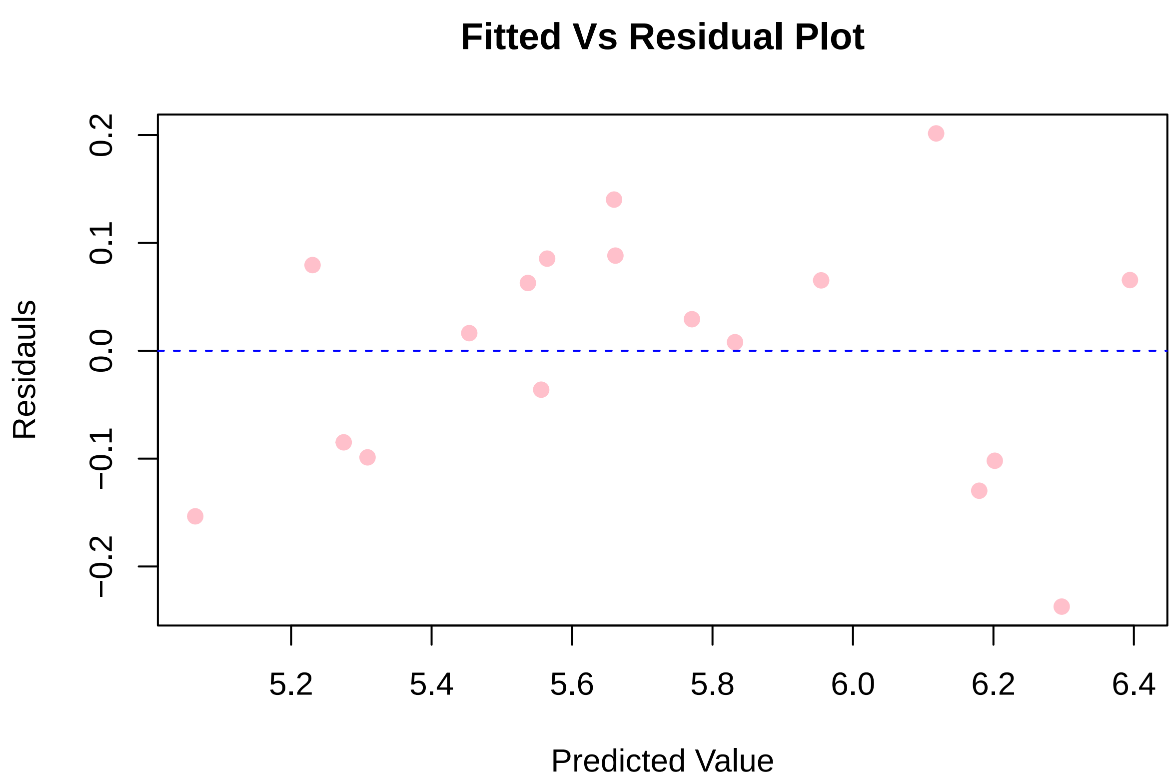
From the above fit we can say that all the coefficients are significant. and the Adjusted R squared is 0.91 and residual sum of square is 0.1213

Residual Plot

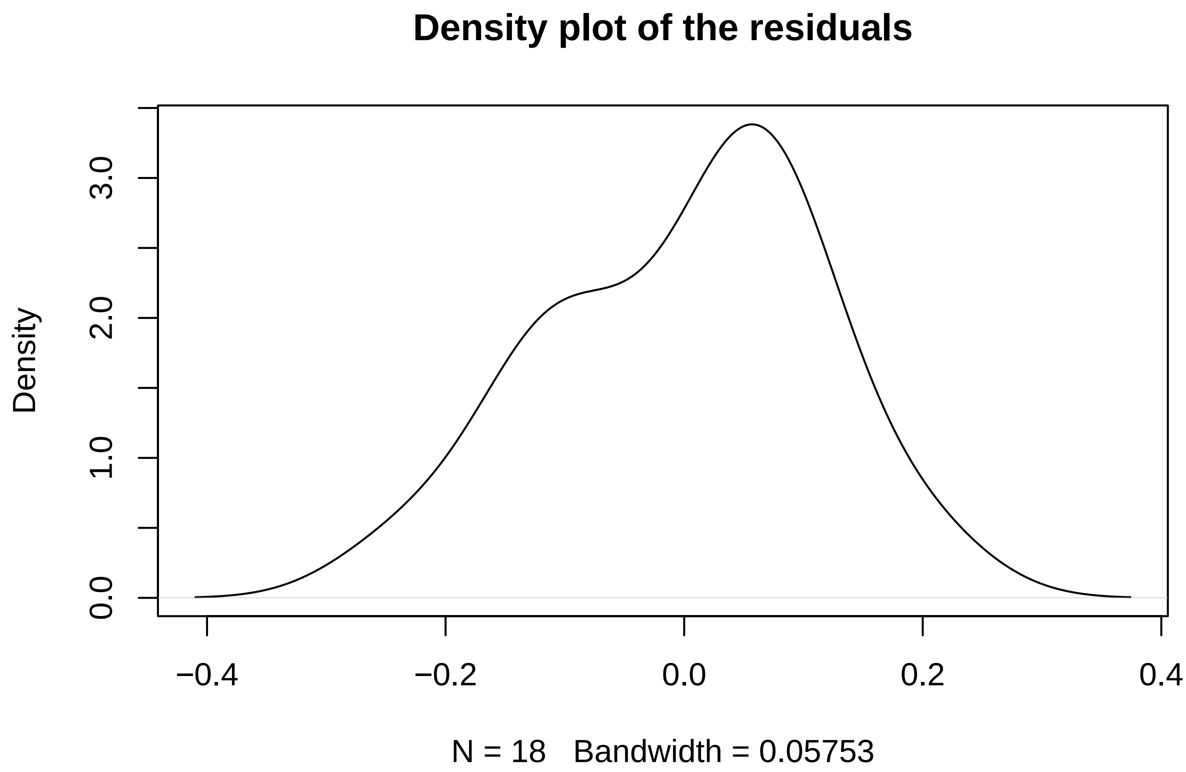
```r
res<- resid(model3)

### Residual sum square

RSS <- sum(res^2)
RSS
```

```
## [1] 0.2207236
```

```r
plot(fitted(model3),res, pch=19,col='pink',
     xlab = "Predicted Value",ylab = "Residauls",
     main = "Fitted Vs Residual Plot")

abline(h=0, col='blue',lty=2)
```




```r
plot(density(res), main = "Density plot of the residuals")
```



The variability of the model is not stabilized and also the errors doesn't follows the assumption of normality so we need transformation of response variable.

Transformation of Variable

```
## Geometric Mean of Y
y_g <- exp(mean(log(df2$y)))
```

```
lambda <- c(-2,-1, -0.5,0, 0.5,2,3)
rss <- c()
a<- c()

for(i in 1:length(lambda)){

  if(lambda[i]!= 0){
    m <-lm( ((y^lambda[i])/y_g) ~ x1 + x2 , data = df2 )
  }
  if(lambda[i]==0){
    m <-lm( log(y) ~ x1 + x2 , data = df2 )
  }
  rss <- c(rss, sum((summary(m)$residuals)^2))
  a<- c(a, summary(m)$adj.r.squared)
}

library(knitr)
```

```
kable(cbind(lambda,"RSS"=round(rss,9),"Adj_R_squared"=a))
```

lambda	RSS	Adj_R_squared
-2.0	0.0000012	0.8812271
-1.0	0.0000080	0.8956760
-0.5	0.0000106	0.9015071
0.0	0.0074078	0.9063771
0.5	0.0003080	0.9102693
2.0	0.8507640	0.9159984
3.0	63.7191012	0.9148843

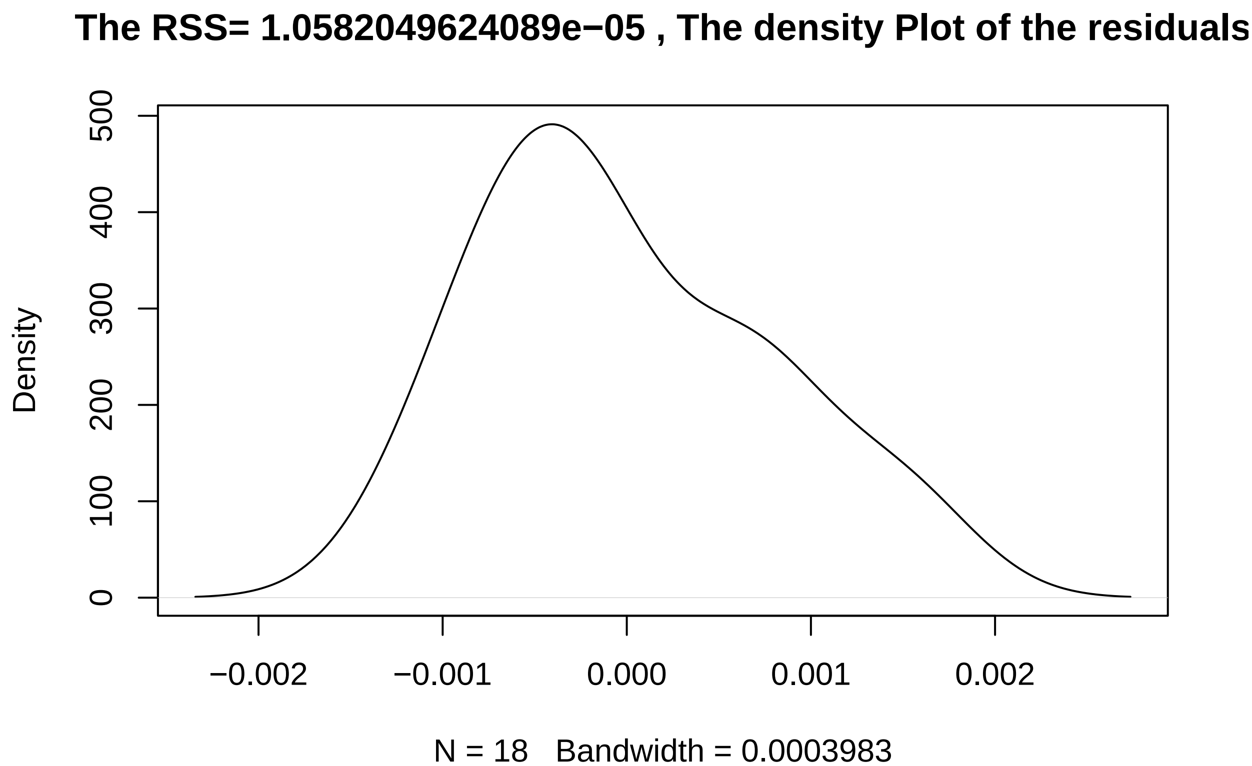
So the model fits well when the $\lambda = -2$, so the model will be more precise if we take $\lambda = 2$ and transform the response variable.

THe model with lambda =-0.5

```
model4 <-  m <-lm( ((y^(-0.5))/y_g) ~ x1 + x2 , data = df2 )
summary(model4)
```

```
##
## Call:
## lm(formula = ((y^(-0.5))/y_g) ~ x1 + x2, data = df2)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.0011481 -0.0005297 -0.0001685  0.0006283  0.0015403
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.072e-02  1.549e-03  58.553  < 2e-16 ***
## x1          -5.387e-04  4.415e-05 -12.200 3.45e-09 ***
## x2          -2.018e-04  3.461e-05  -5.832 3.30e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0008399 on 15 degrees of freedom
## Multiple R-squared:  0.9131, Adjusted R-squared:  0.9015
## F-statistic:  78.8 on 2 and 15 DF,  p-value: 1.104e-08
```

```
res2 <- resid(model4)
plot(density(res2), main = paste("The RSS=",sum(res2^2),", The density Plot of the residuals"))
```



Final Model

$$y^* = \widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \widehat{\beta}_2 x_2$$

where

$$y^* = \frac{y^{-1/2}}{Y_g}$$

Here the estimated coefficients are,

$$y^* = (9.072e - 02) + (-5.387e - 04) \times x_1 + (-2.018e - 04) \times x_2$$

This is the final model.

Here is a big question........ which lambda should we choose