REGRESSION BASED ANALYSIS OF WORLD HAPPINESS REPORT 2021

SUBMITTED BY-

NISTHA SHAH (201358)

ASHISH KUMAR (201283)

SHWETANK SINGH (201422)

RAHUL KUMAR SINGH (201378)

AKASH KUMAR SHARMA (201261)

(Under the guidance of Dr. Sharmistha Mitra)

Acknowledgement

We would like to express our heartfelt gratitude to **Dr. Sharmishtha Mitra, Department of Mathematics and Statistics, IIT Kanpur** for assigning this project to us. It has been a great learning experience and has also provided us with a practical insight of the theoretical knowledge gathered during the course lectures.

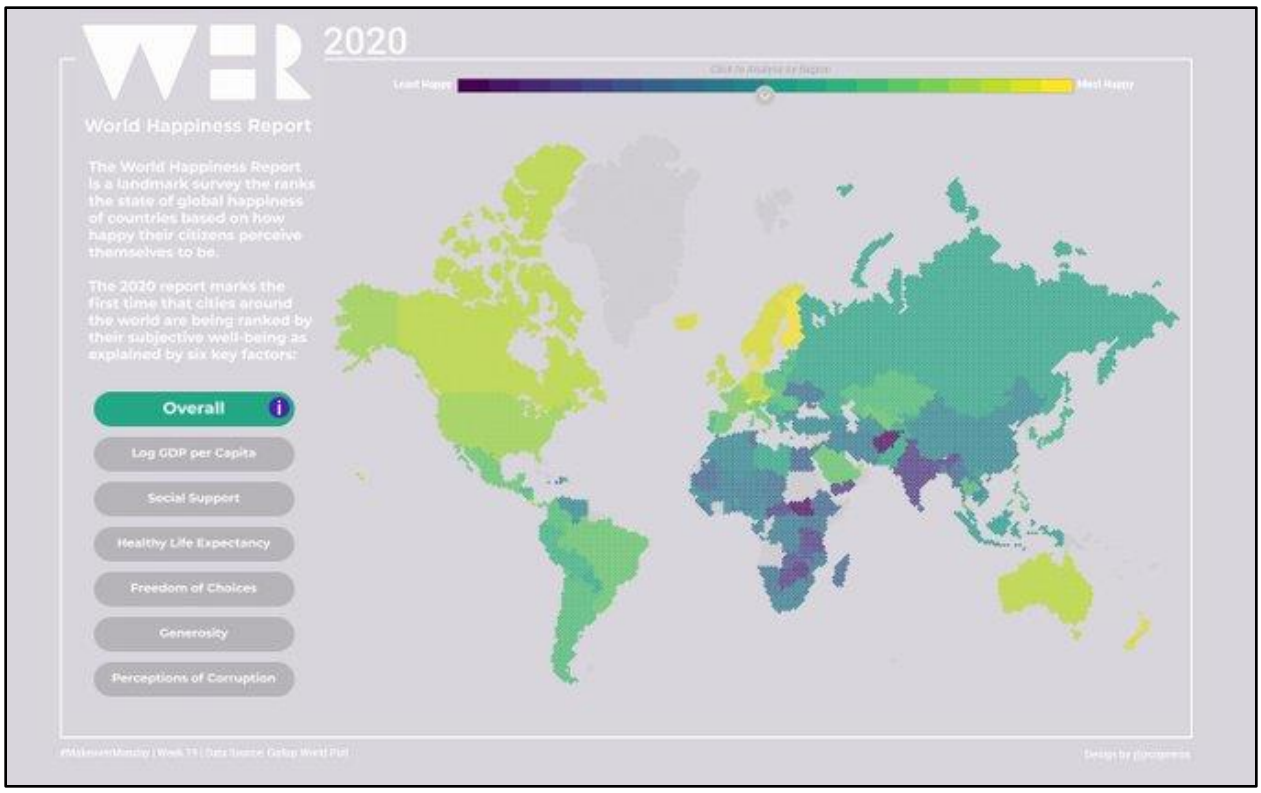
We would also like to thank our seniors for their extensive support throughout the session. Their constant encouragement has enabled us to complete the project within the stipulated time- period.

Contents

Index	Title	Page Number
1	Introduction	1-3
2	Objective	4
3	About the Data	5-6
4	Model & Assumptions	7
5	Parameter estimates & Summary	8
6	Checking Skewness & Outliers	9-13
7	Detection & Removal of Outliers	13-15
8	Checking Normality Assumptions	16-17
9	Checking Multicollinearity	18
10	Multicollinearity Diagnostics	19-24
11	Heteroscedasticity	25-30
12	Predictive Power	31
13	Ridge Regression Vs PCR Model	31
14	Final Model & Summary	32
15	Plot of Predicted Vs Original Ladder Score	33-34
16	R Codes	35-46
17	Bibliography	47

Introduction

- **Happiness: What and how?**
- **Different Countries and Prediction of their hypothetical Score.**



What is happiness and how it can be derived?

Happiness is a fuzzy concept and can mean many different things to many people. The question is as old as mankind itself. Still, there is no consensus however about what happiness is. Part of the challenge of science of happiness is to identify different concept of happiness, and where applicable, split them into their components. And that is why measuring happiness is a tough task. Related concept to it, are subjective well-being quality of life and flourishing.

The **2020 World Happiness Report** stated that in subjective well-being measures, the primary distinction is between cognitive life evaluations and emotional reports. Happinessis used in both life evaluation, as in "**How happy are you with your life as a whole?**", and in emotional reports, as in **"How happy are you now?**" We have distinguished determinants, falling under five broad categories, at play answering this question concerning subjective well-being. These include socio-demographic factors,

health/well being factors, institutional determinants, economic factors and other miscellaneous factors (that do not belong to above four). These chosen indicators are described in detail as the report progresses.

Why we are interested in building a happiness model ?

We have enumerated the following reason to answer this question*:*

- Because there is a growing global interest in including **happiness and subjective well-being** as a primary indicator of the quantity of human development.

- Many governments, communities and organizations are using happiness data, and theresults of subjective well-being research, to **make policies** in line with what really matters to people and that support better lives.

- We're interested in the relationship between overall happiness and a wide variety of **cultural, geographical and macroeconomic factors**.

- Last but not the least, we're curious to know which elements should we focus on, to keep ourselves happy post graduation (once we enter the real big bad world).

Objective

- The goal is to model happiness over various factors like, GDP per capita, Social Support, health, Freedom, Generosity and Perception of Corruption, subject to data being collected from countries all over the world.

- We want to build a model, such that it can be used to predict global happiness indices for the next year, with maximum precision.

ABOUT THE DATA

RESPONSE VARIABLE:-

LADDER SCORE(y):- It asks respondents to think of a ladder, with the best possible life for them being a 10, and the worst possible life being a 0. They are then asked to rate their own current lives on that 0 to 1.

PREDICTORS:-

LOGGED GDP PER CAPITA(x_1):- Per capita gross domestic product (GDP) is a metric that breaks down a country's economic output per person and is calculated by dividing the GDP of a country by its population.

SOCIAL SUPPORT(x_2):- is the national average of the binary responses (0=no, 1=yes) to the Gallup World Poll (GWP) question. "If you were in trouble, do you have relatives or friends you can count on to help you whenever you need them, or not?"

HEALTHY LIFE EXPECTANCY(x_3):-The time series of healthy life expectancy at birth are constructed based on data from the World Health Organization (WHO) Global World Health Observatory data repository, with data available for 2010, 2015, and 2018. To match this report's sample period, interpolation and extrapolation are used.

FREEDOM TO MAKE LIFE CHOICES(x_4):- Is the national average of binary responses to the GWP question, "Are you satisfied or dissatisfied with your freedom to choose what you do with your life?"

GENEROSITY(x_5):- Is the residual of regressing the national average of GWP responses to the question. "Have you donated money to a charity in the past month?" on GDP per capita.

PERCEPTION OF CORRUPTION (x_6):- Are the average of binary answers to two GWP questions; “Is corruption widespread throughout the government or not?” and “Is corruption widespread within businesses or not?” Where data for government corruption are missing, the perception of business corruption is used as the overall corruption-perception measure.

Observations:- We have taken the data of year 2020. There are 149 observation in total.

Source:-

We have collected data from following source:

https://worldhappiness.report/ed/2021/

Model & Assumptions

Model:- The linear Regression model is given by:

$$Y_i=\beta_0+\sum_{j=1}^{6} x_{ij}\beta_j+\varepsilon_i \quad ; \quad i=1(1)149$$

Assumptions:-

The main assumption in linear Regression model is following:

(i) ε_i's are Normally distributed.

(ii) $E(\varepsilon_i)= 0$ and $V(\varepsilon_i)=0$ for all i=1(1)n i.e. error are homosedastic.

(iii) $(\varepsilon_1,\varepsilon_2,\ldots\ldots\ldots\ldots,\varepsilon_n)$ are independently distributed i.e. error are uncorrelated.

(iv) $Cov(x_i , x_j)=0$ for all i≠j i.e. Model is free from multicollinearity problem.

Parameter Estimate and Summary

Parameter estimate:- The estimated value of ordinary list square estimation method are given below:

Coefficients:

	Estimate	Std. Error	t-value	Pr(>|t|)	Significance
(Intercept)	-2.23722	0.63049	-3.548	0.000526	***
x1	0.27953	0.08684	3.219	0.001595	**
x2	2.47621	0.66822	3.706	0.000301	***
x3	0.03031	0.01333	2.274	0.024494	*
x4	2.01046	0.49480	4.063	7.98e-05	***
x5	0.36438	0.32121	1.134	0.258541	
x6	-0.60509	0.29051	-2.083	0.039058	*

Significance Codes: '***'0.001 '**'0.01 '*'0.05 '.'0.1 ' '1

Summary of the model:-

Residual standard error: 0.5417 (with 142 df).

Model R-squared: 0.7558

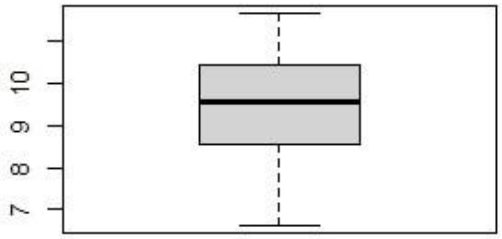
Adjusted R-squared: 0.7455

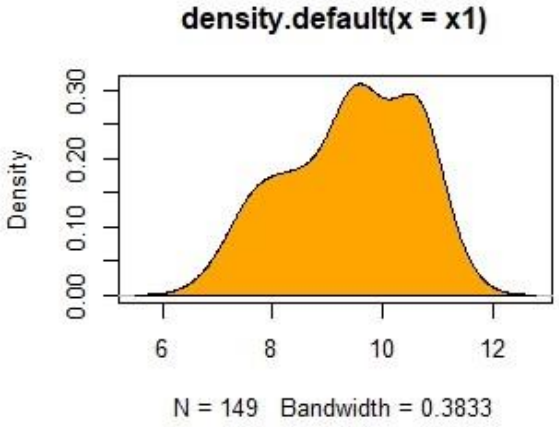
p-value: < 2.2e-16

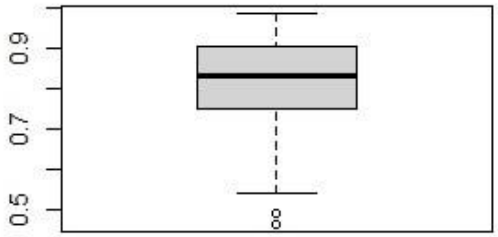
F-statistic: 73.27 (on 6 & 142 df).

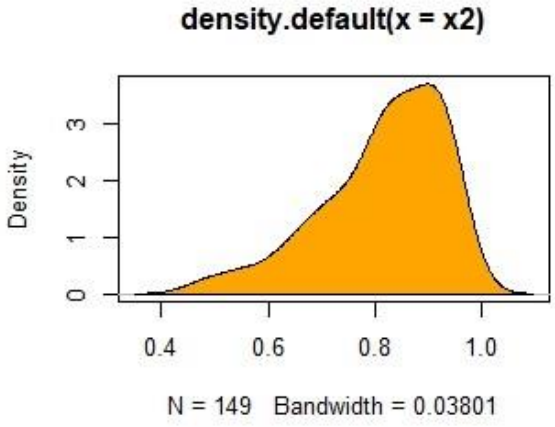
CHECKING SKEWNESS IN THE DATA

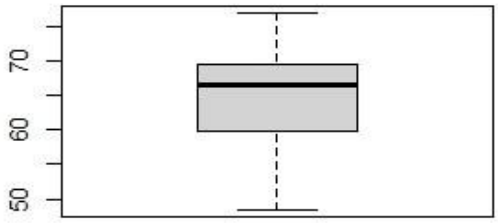
We have 6 regressors after removing multicollinearity, and some of the regressors are found to be highly skewed. The Boxplot of regressors and there corresponding densities are:

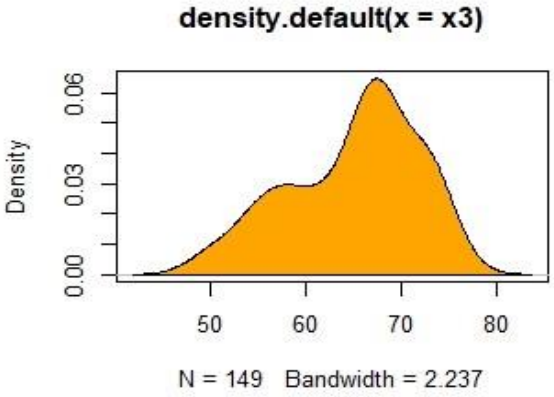


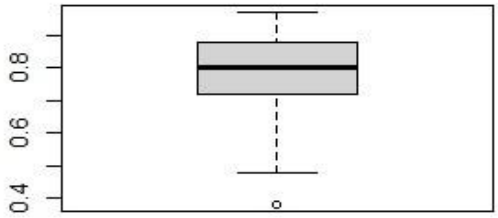


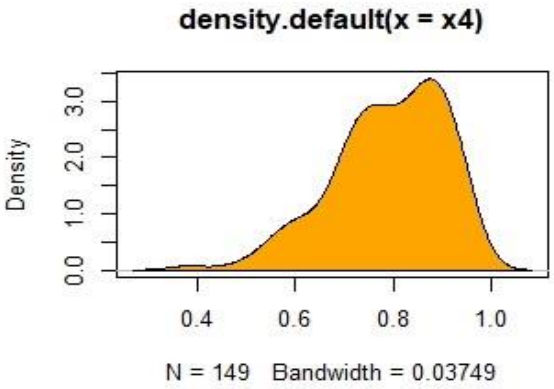


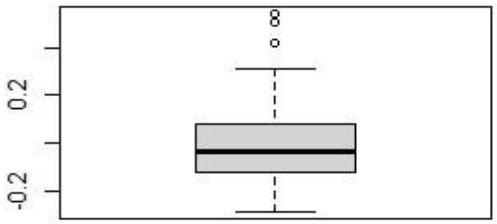


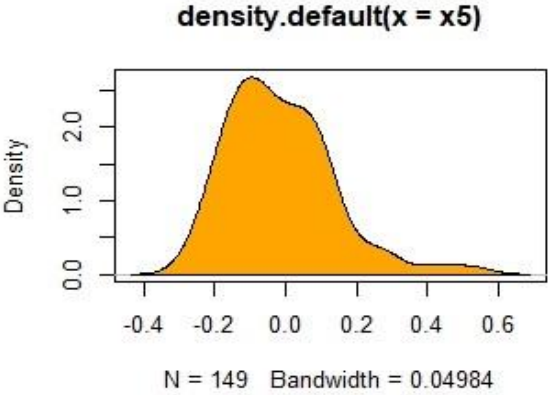

50
60
70

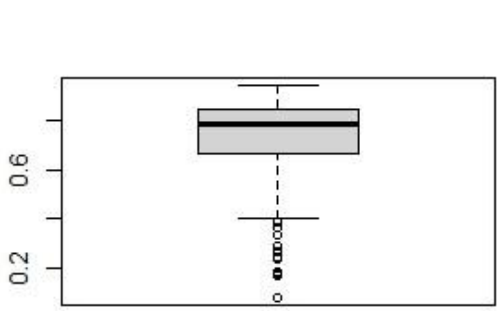

density.default(x = x3)
Density
0.00
0.03
0.06
50
60
70
80
N = 149 Bandwidth = 2.237

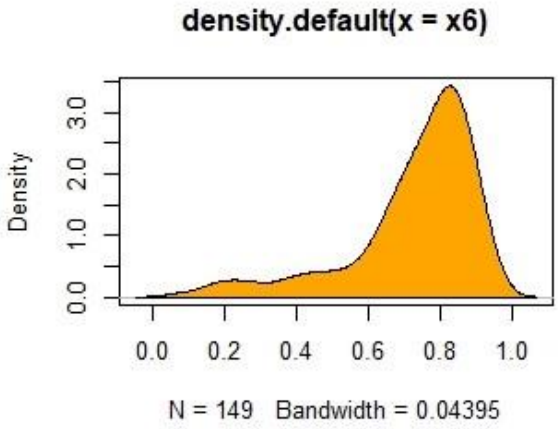

0.4
0.6
0.8


density.default(x = x4)
Density
0.0
1.0
2.0
3.0
0.4
0.6
0.8
1.0
N = 149 Bandwidth = 0.03749


-0.2
0.2


density.default(x = x5)
Density
0.0
1.0
2.0
-0.4
-0.2
0.0
0.2
0.4
0.6
N = 149 Bandwidth = 0.04984

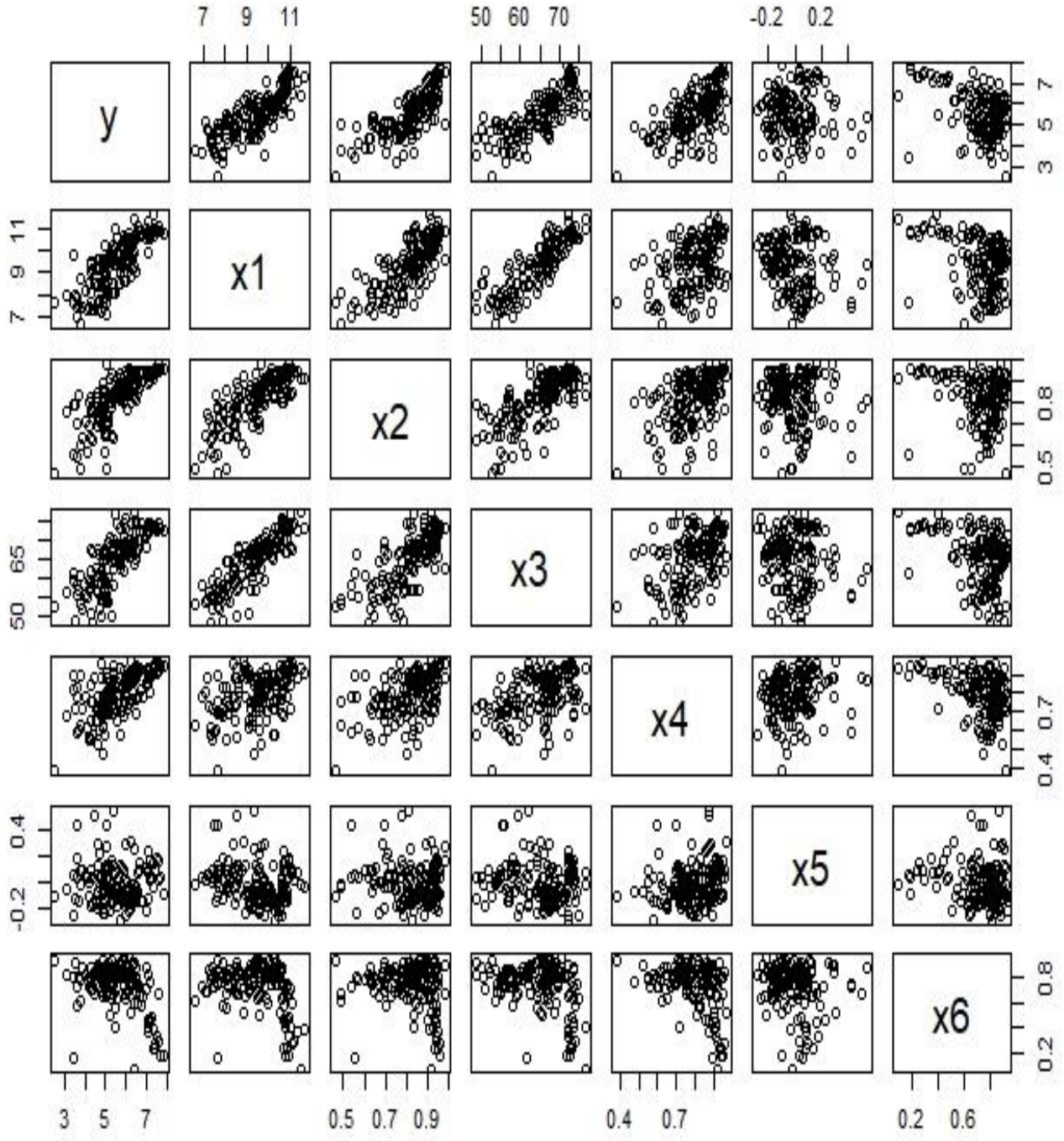




We found that the none of the regressors are highly skewed. So we haven't use any transformation.

Skewness of x_1 , x_2 , x_3 , x_4, x_5 and x_6 are respectively -0.3449691, -0.9190126, -0.511535, -0.7396695, 0.9897123, -1.545846

Observation from scatter plot matrix:-



i. There is linear relationship between 'Logged GDP per capita' (X1) & 'Social support' (X2).
ii. There is linear relationship between 'Logged GDP per capita' (X1) & 'Healthy life expectancies' (X3).

Detection and Removal of Outliers:-

In statistics, an outlier is defined as an observation which stands far away from the most of other observations. Often an outlier is present due to the measurements error. Therefore, one of the most important task in data analysis is to identify and (if is necessary) to remove the outliers.

Outliers can occur by chance in any distribution, but they often indicate either measurement error or that the population has a heavy-tailed distribution. In the former case one wishes to discard them or use statistics that are robust to outliers, while in the latter case they indicate that the distribution has high skewness and that one should be very cautious in using tools or intuitions that assume a normal distribution.

There is no rigid mathematical definition of what constitutes an outlier; determining whether or not an observation is an outlier is ultimately a subjective exercise. There are various methods of outlier detection.

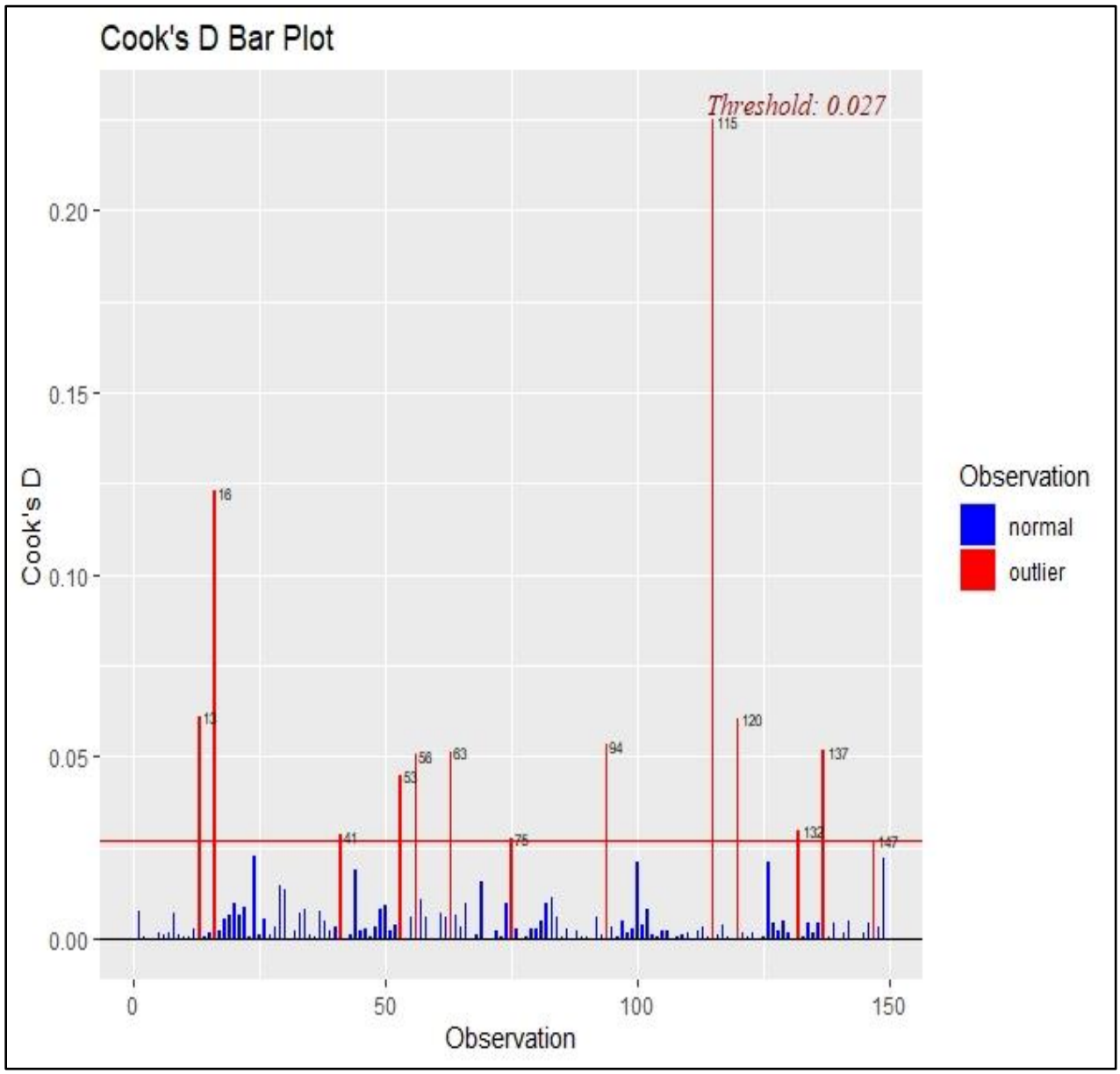
Cooks Distance: Cook's distance, Di, is used in Regression Analysis to find the influential outliers in a set of predictor variables In other words, it's a way to identify points that negatively affect your regression model. The measurement is a combination of each observation's leverage and residual values; the higher the leverage and residuals, the higher the Cook's distance.

Technically, Cook's D is calculated by removing the ith data point from the model an recalculating the regression. It summarizes how much all the values in the regression model change when the ith observation is removed.

The formula for cook's distance is:-

$$D_i = \frac{\sum_{j=1}^{n}(\hat{Y}_j - \hat{Y}_{j(i)})^2}{(p+1)\,\hat{\sigma}^2}$$

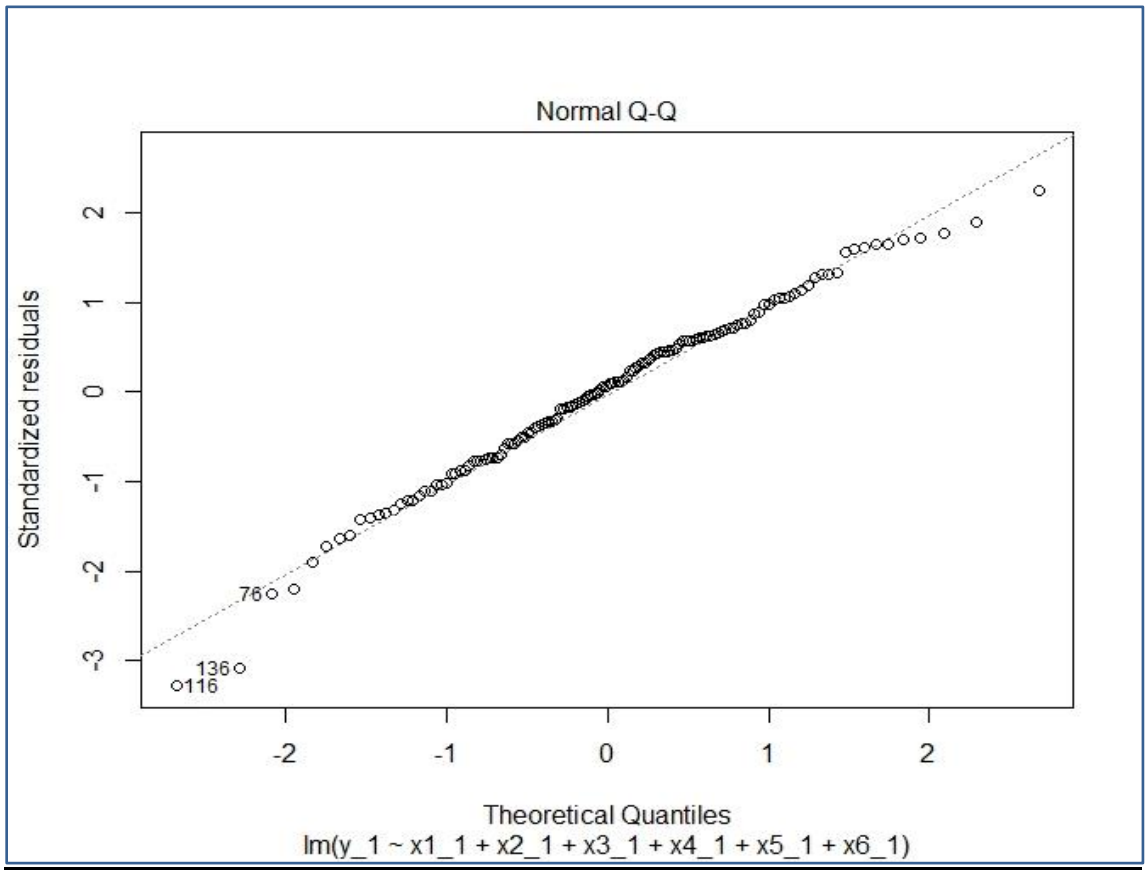
Where $Y_j(i)$ is the fitted response value obtained when excluding.



Checking for Normality assumptions :-

- **Graphical method**:- First we've checked the Q-Q plot of the regressand(Y)

Normal Q-Q plot:-



Since the data points are approximately lie on the 45˚ diagonal line we can assume normality for Y and ε.

- **Shapiro-Wilk test:-**

  H_0: Errors are normally distributed ag. H_1:H_0 is not true

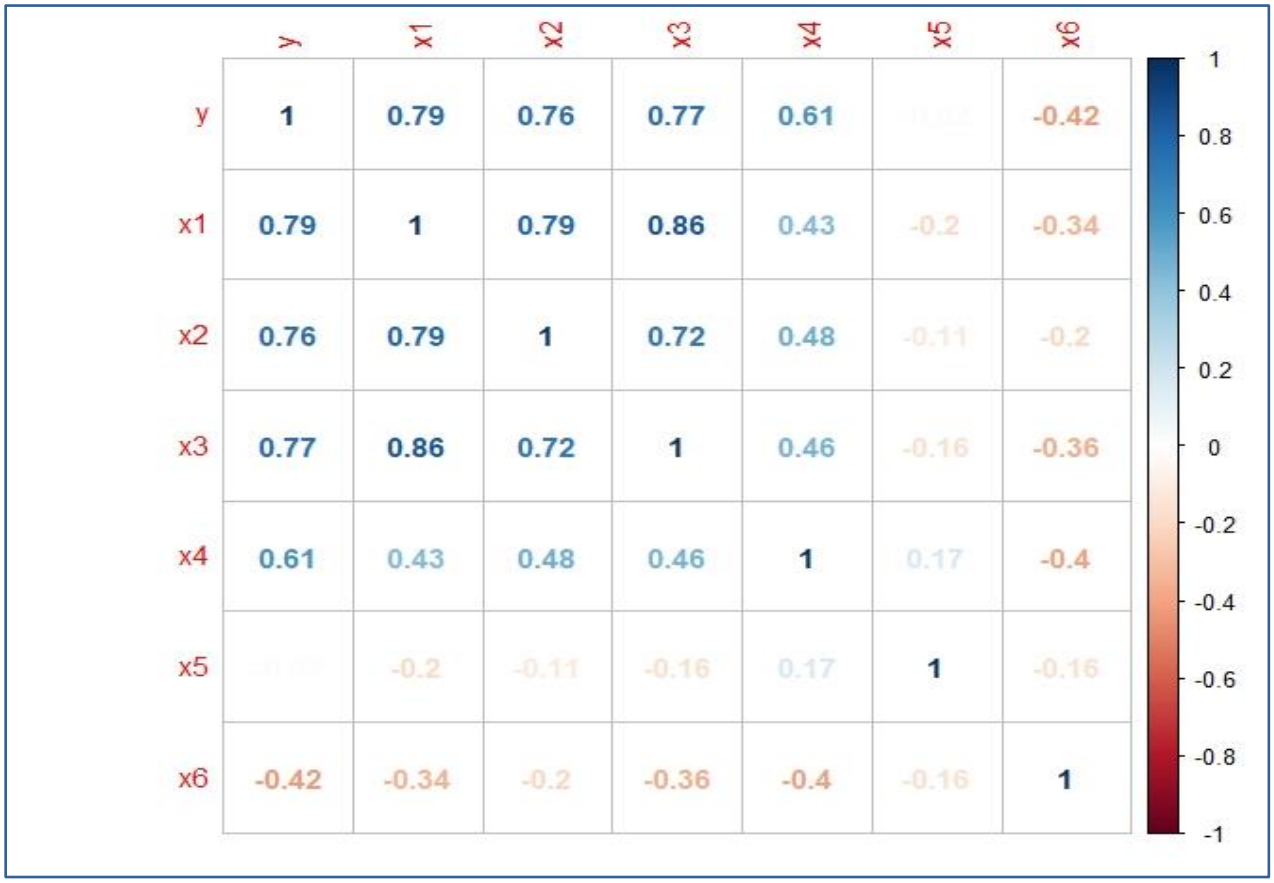
  Test statistic, **W = 0.98465, p-value = 0.1315**>0.05(α)

  Since, the null hypothesis (H_0) is accepted at 5% level of significance, wecan confirm that Y and hence the errors follow a normal distribution.

Check Multicollinearity

Correlation Matrix

It is correlation matrix which gives correlation between response variable and regressor.



This matrix tells that there are regressors that are linearly related to each other. Hence gives an indication that there is multicollinearity present in some group of variables.

Multicollinearity Diagnostics

- **Variance Inflation Proportion(VIP) Method:** This method is used to identify regressors that are involved in multicollinearity.

Variance Inflation Factor (VIF):

$$VIF_j = \frac{1}{1-R_j^2}$$

Where, R_j^2 is the coefficient of multiple determination obtained from regressing x_j on the other remaining regressor variables.
VIF value greater than 4 indicate multicollinearity and large value of multicollinearity leads to poor estimates of regression coefficient. VIF's of each regressors are as follows:

Variable	VIF
x_1	5.104890
x_2	2.972200
x_3	4.099348
x_4	1.585807
x_5	1.180982
x_6	1.367122

We see that, VIF corresponding to regressors 'logged GDP' (x_1) and 'healthy life' (x_3) are greater than 4 which indicates that these two variables are involved in multicollinearity.

- **Ridge Regression and Principal Component Regression(PCR):**

  Since we have only six regressors in our data and we consider all of them important to predict ladder score. So instead of eliminating them we will see effect of the each regressor on the response. For this, we've used Ridge regression and principal component regression(PCR) to reduce multicollinearity. Then we will compare the efficiency of these two methods.

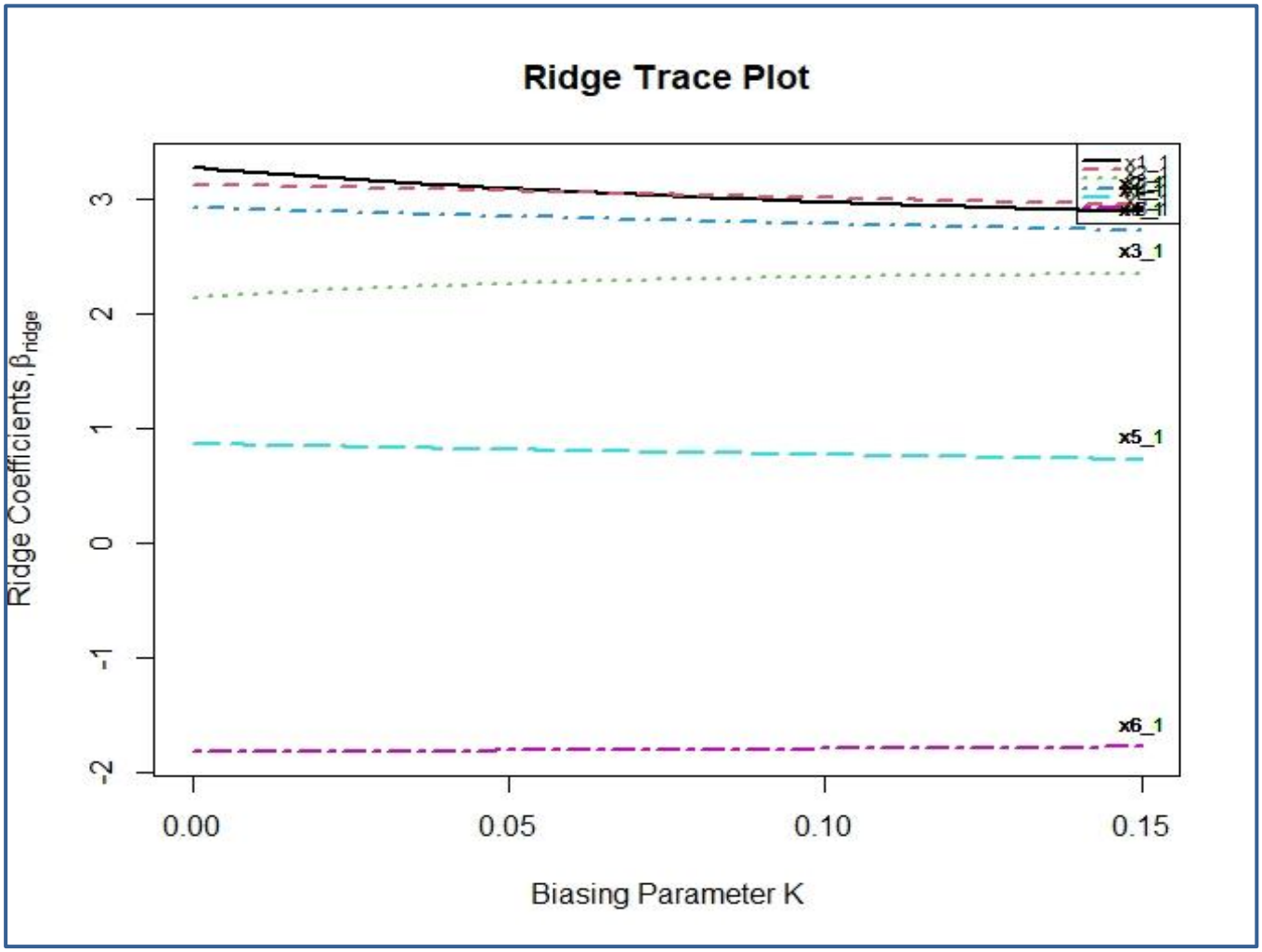
Ridge Regression:

The ridge function for fitting ridge regression is given by:

$$R(\beta) = (Y-X\beta)' (Y-X\beta) + \lambda \beta' \beta$$

We have chosen the value of λ by two methods. First one is Ridge Trace Plot, this method is graphical method and appropriate value of λ is chosen from the trace plot and second one is Hoerl & Kennard iterative procedure.

- **Ridge Trace Plot:**



From the above plot we observe that appropriate value λ should be around λ=0.03 so that all the estimates of parameters stabilize.

- **Iterative procedure:** the estimate of λ obtained by this method is 0.00297.

Observations:

- We check that after fitting ridge regression model the VIFs have decreased significantly.
  New VIF's of each regressors are as follows:

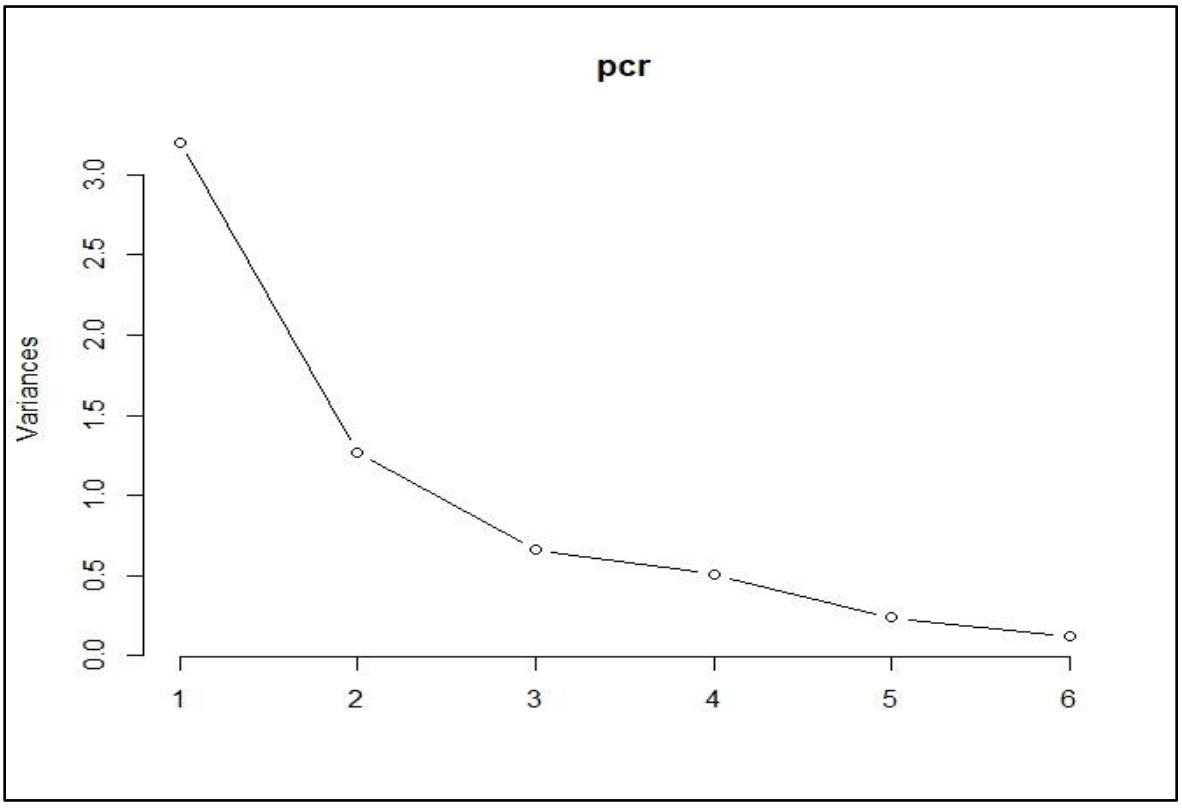
Variable	VIF
x_1	3.50804
x_2	2.57029
x_3	2.99295
x_4	1.47156
x_5	1.09597
x_6	1.20986

- The adjusted R-sq is 80.12%

Principal Component Regression:

```
Importance of components:
                          PC1    PC2    PC3    PC4     PC5     PC6
Standard deviation     1.790 1.1264 0.8125 0.7129 0.48785 0.34762
Proportion of Variance 0.534 0.2115 0.1100 0.0847 0.03967 0.02014
Cumulative Proportion  0.534 0.7455 0.8555 0.9402 0.97986 1.00000
```

We can explain 94.02% of variation using the first 4 principal components. After that the increase in proportion of variance is not so significant. So, we choose 4 principal components from the 6. Also, it can be seen from the PCR scree plot which plots the cumulative variances corresponding to each principal component.



Observations:

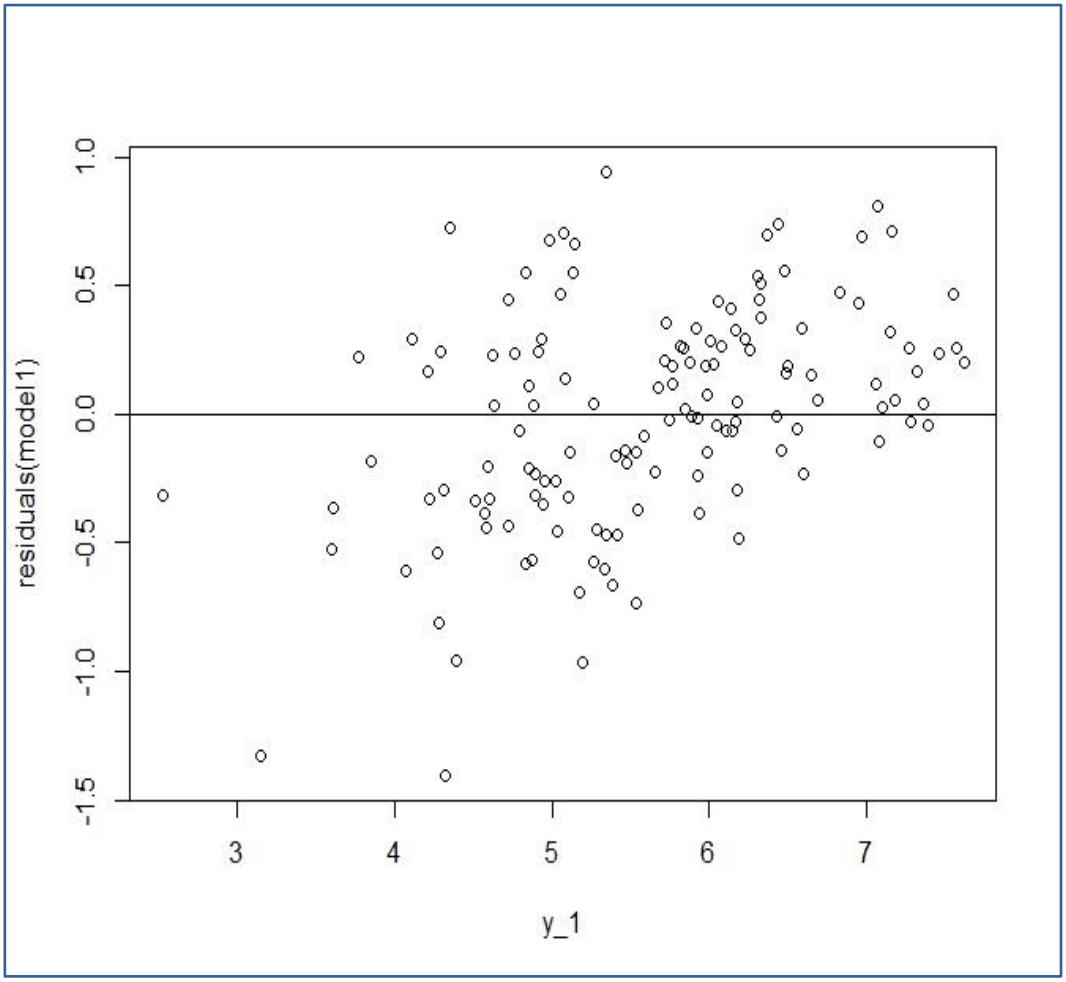
- After doing PCR, we check that the VIFs corresponding to each of the 3 principal components become 1. So, we have successfully handled the problem of multicollinearity in our model.
- The adjusted R-sq is 81.82%.

Now, we analyze the two models separately and check which one is better.

Heteroscedasticity

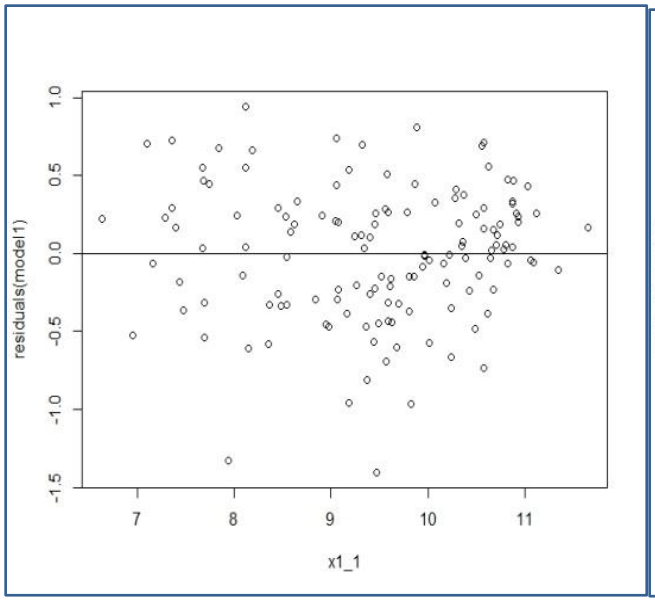
In regression analysis it is required that the errors are homoscedastic. The errors have mean zero and constant variance and are uncorrelated. So Y should also be homoscedastic in nature**.**

1. **Using Ridge Regression Model:-**
   a) **Graphical method:-** We've first checked the residual plot vs fitted values of regressand (Y) for any unusual pattern.

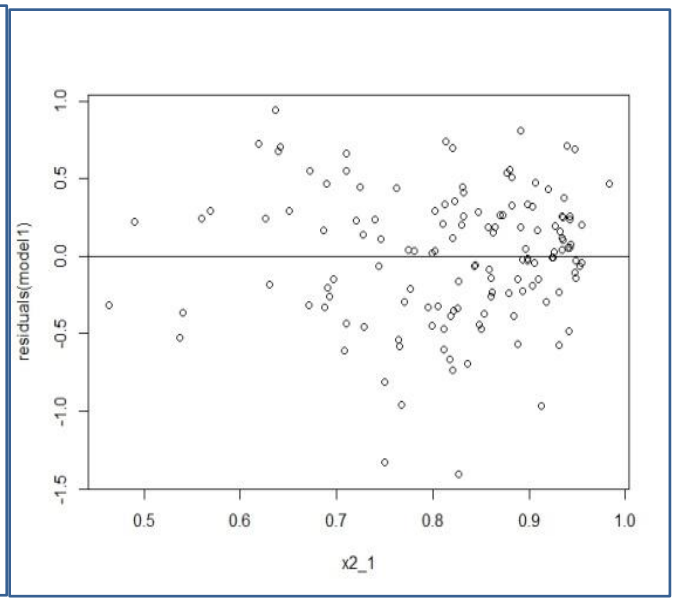


It shows a slight **U shaped** pattern. Next we've found out the residual plots vs. each of the predictors to check which of these are responsible for this pattern.

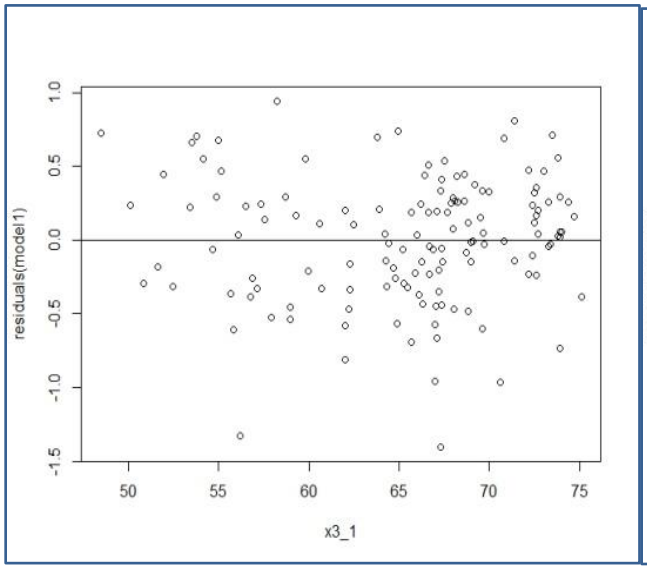
Residual plot vs x_1



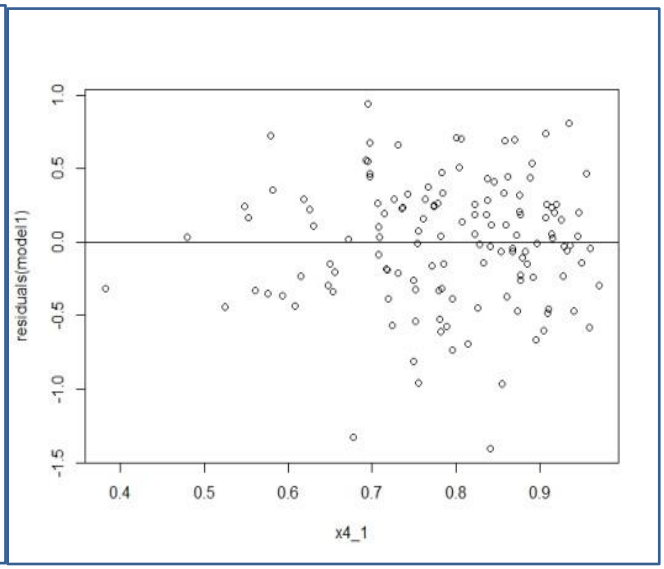
Residual plot vs x_2



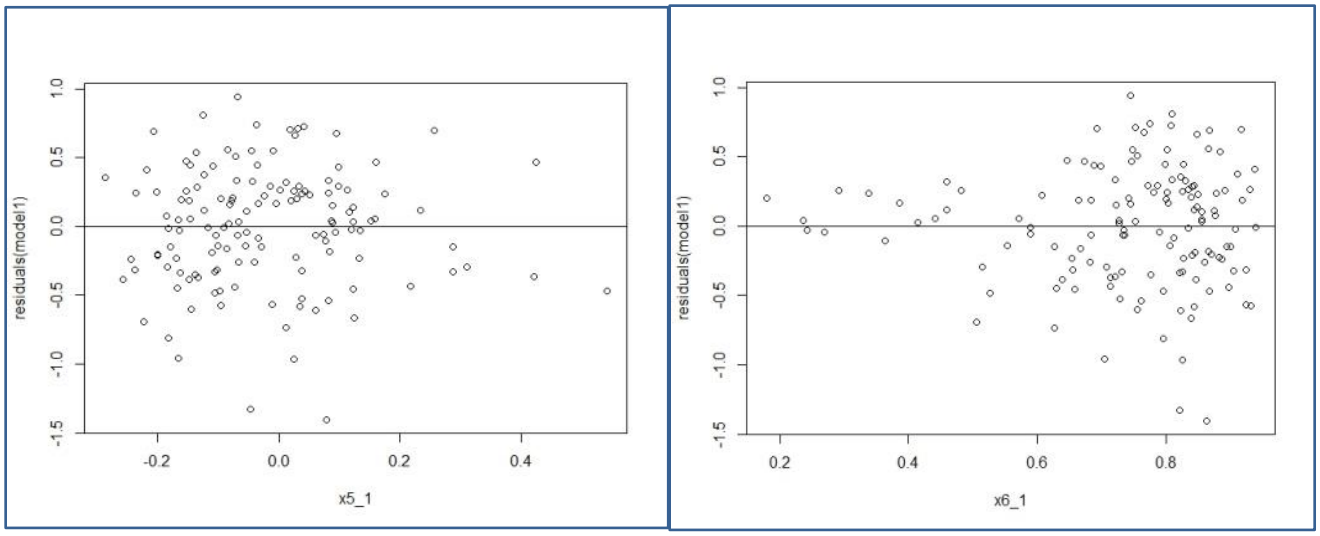
Residual plot vs x_3



Residual plot vs x_4



Residual plot vs x_5 Residual plot vs x_6



Observation:-

The plots vs x_4,x_5 & x_6 show some unusual pattern which needs to be checked further.

b). Glejser's test:- This test checks which of the predictors are responsible for heteroscedasticity and the form of heteroscedasticity. It tests for H_0:c_1=0 in the model:

$$|e_i| = c_0 + c_1 x_{ij}^{c_2} \qquad \text{i=1(1)149, j=1(1)6}$$

We check the value of R-sq for each of these cases taking c_2 = -1, -½, ½, 1.

	x	1/x	$\sqrt{X}$	$1/\sqrt{X}$
X_4	0.004883	0.001402	0.003903	0.00213
X_5	8.104e-08	0.0008338	-	-
X_6	0.04368	0.04796	0.04748	0.05012

c). Goldfeld Quandt test:-

It's a confirmatory test for determining the exact form of heteroscedasticity. We perform this test for themost significant forms of the variables x_4 and x_6.

The p-values are respectively 0.09882 & 0.1265, since both are greater than 0.05 so this indicates that heteroscedasticity is not significant.

d). Breusch–Pagan test:-

In statistics, the Breusch–Pagan test (named after Trevor Breusch and Adrian Pagan) is used to test for heteroscedasticity in a linear regression model. It tests whether the estimated variance of the residuals from a regression are dependent on the values of the independent variables. In that case, we have heteroscedasticity in our model.

Ho: Residuals are homoscedastic

Against

H1: Ho is not true.

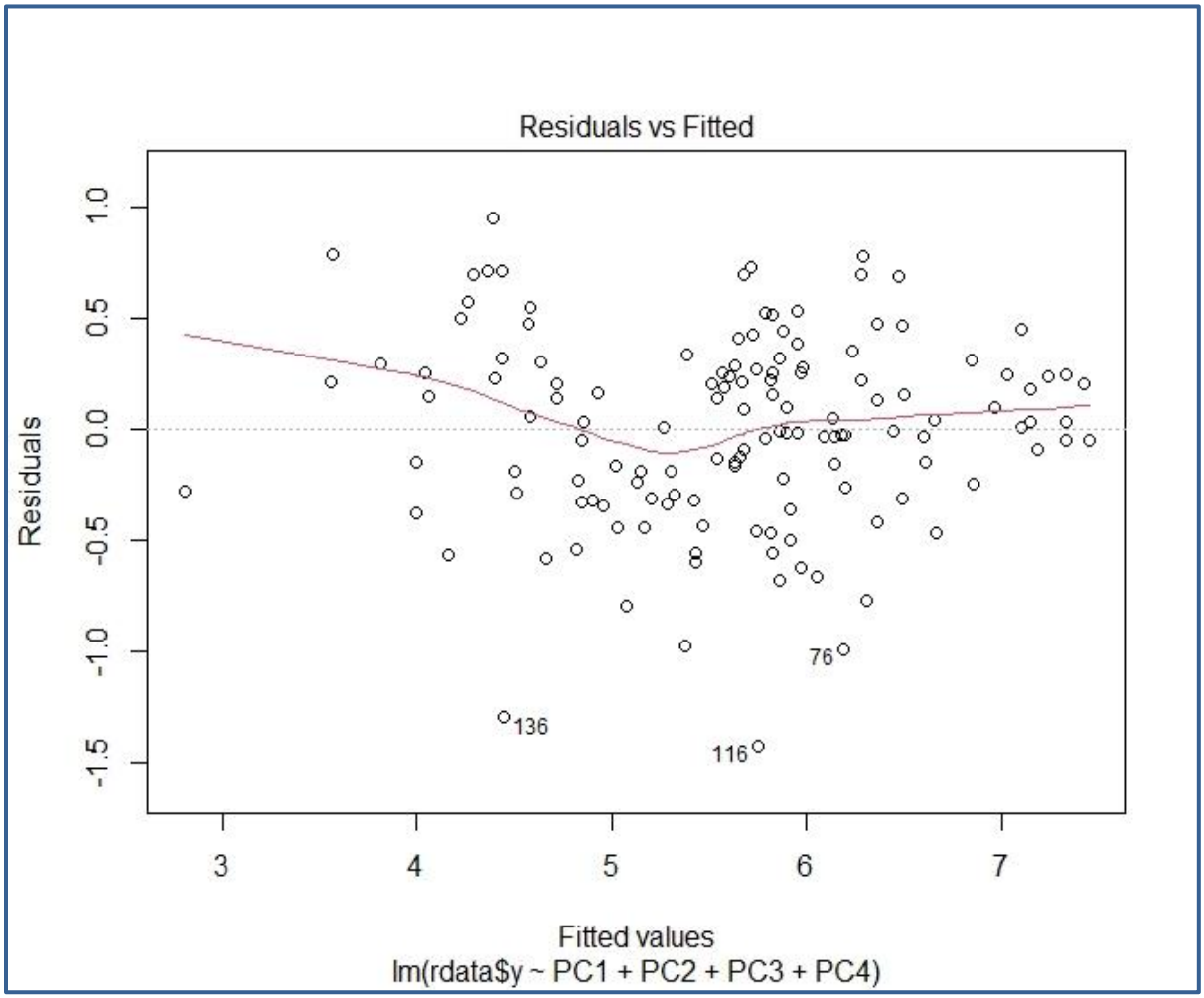
```
        studentized Breusch-Pagan test

data:  model2
BP = 6.7128, df = 4, p-value = 0.1519
```

Since the p-value is more than α = 0.05, we accept the null hypothesis that the data is homoscedastic.

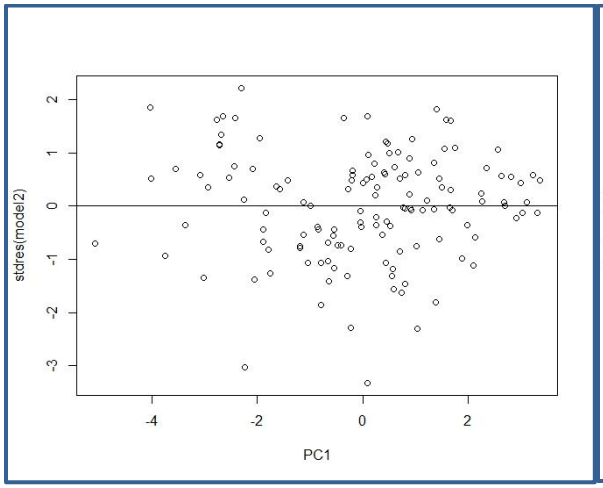
Using PCR Method:-

Graphical method: We've first checked the residual plot vs fitted values of regressand (y) for any unusual pattern.

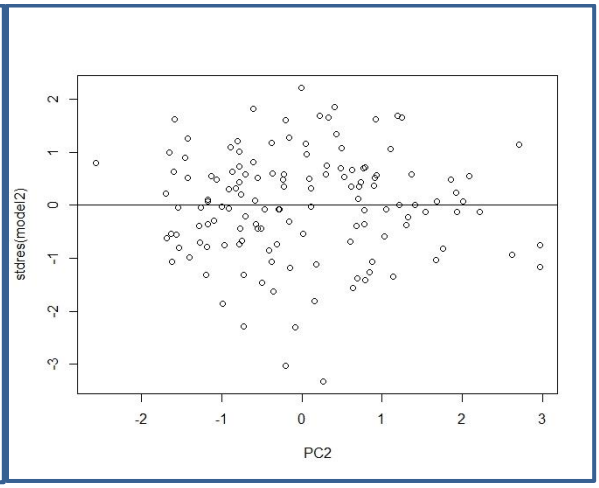


There is a slight **S** shaped pattern in the plot. So, we've found out the residual plots vs. each of the predictors to check which of these are responsible for this pattern.

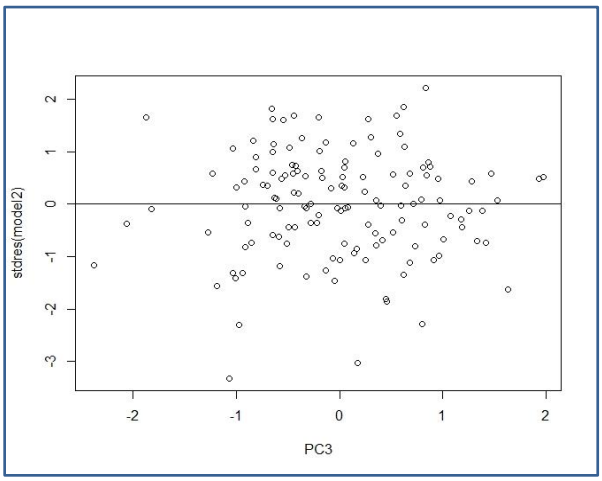
Residual plot vs PC1



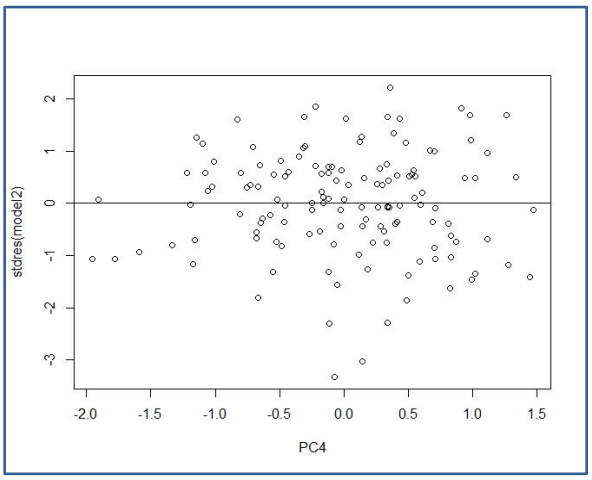
Residual plot vs PC2



Residual plot vs PC1



Residual plot vs PC2



There is no significant pattern in these plots. So, we don't go for any further tests.

Predictive Power

Since we've obtained our final model through two methods, now we proceed to check the usefulness of these two models for predicting purpose. Hence, we compute R-sq predicted using the PRESS statistic.

For Ridge Regression Model, $R^2_{pred} = 80.86\%$

For PCR Model, $R^2_{pred} = 81.10\%$

We can conclude that these models can be reliably used for predicting production.

Ridge Regression vs PCR Model

I. The Ridge Regression Model shows a slight heteroscedasticity in residual plots whereas in the PCR Model , there is no significant pattern in the residual plots. Though ultimately both models can be treated as homoscedastic.

II. The PCR Model has a slightly better predictive power than Ridge Regression Model.

Final Model and Summary

Model after Ridge Regression:

Y = -1.5781 + 0.2390 x1 + 2.4325 x2 + 0.0298 x3 + 2.1707 x4 + 0.4991 x5 - 0.9676 x6

Summary:

```
Ridge Summary
         R2     adj-R2   DF ridge           F         AIC         BIC
    0.80850    0.80120    5.56065  101.71921 -222.32766  461.98966
Ridge minimum MSE= 2.440751 at K= 0.02976505
P-value for F-test ( 5.56065 , 130.0569 ) = 3.242863e-45
```

Clearly we can conclude that 80.12% of the total variability in the Ladder score is explained by our regression model. Since the p-value of F-statistic is 3.242863e-45, which is less than 0.05, we reject the null hypothesis and interpret that the regression model is significant.

Model after PCR:

Y = 5.612941 + 0.509072 PC1 + 0.112711 PC2 - 0.057716 PC3 - 0.001522 PC4

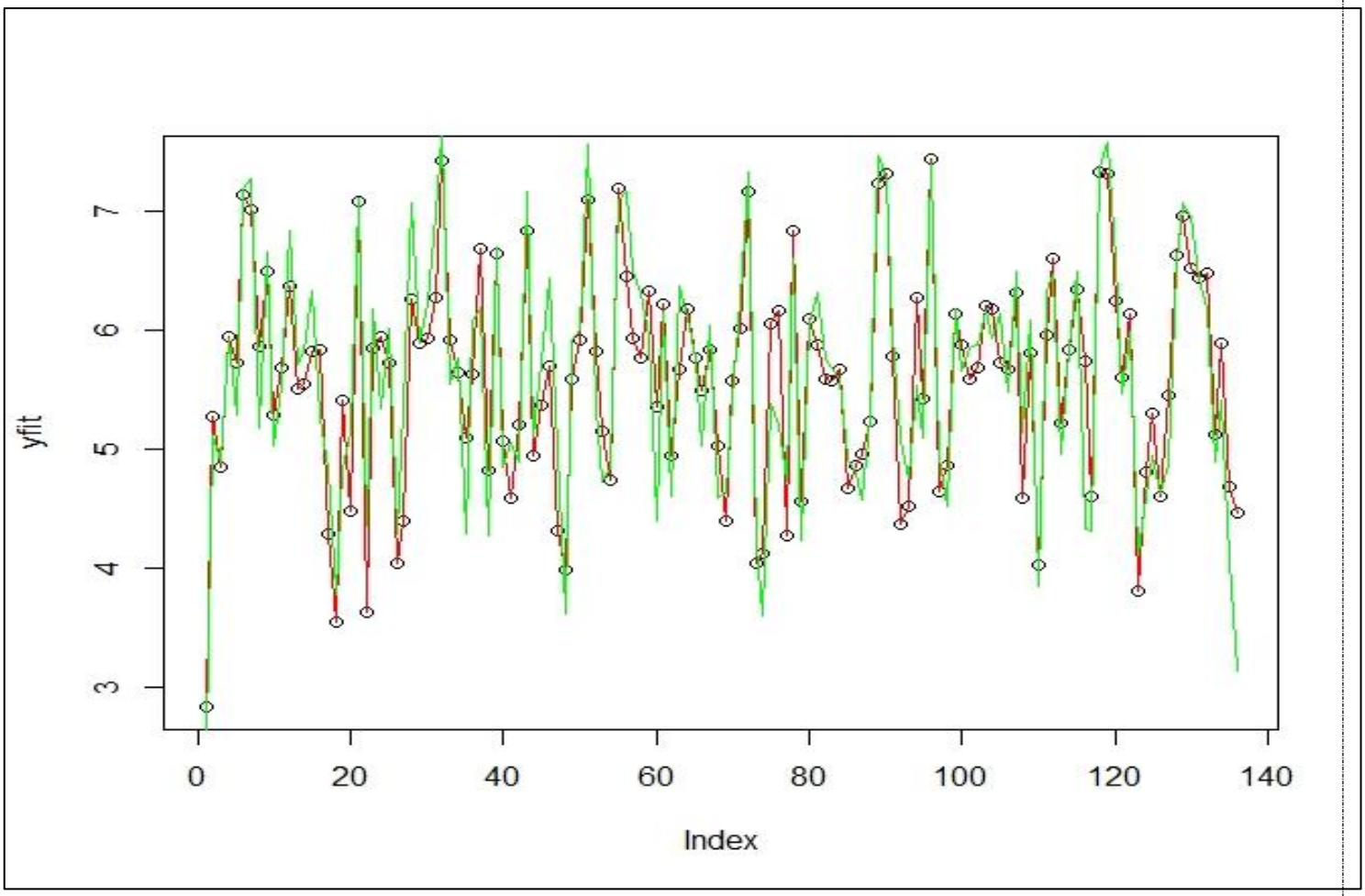
Summary:

```
Residual standard error: 0.4328 on 131 degrees of freedom
Multiple R-squared:  0.8236,    Adjusted R-squared:  0.8182
F-statistic: 152.9 on 4 and 131 DF,  p-value: < 2.2e-16
```

Clearly we can conclude that 81.82% of the total variability in the Ladder score is explained by our regression model. Since the p-value of F-statistic is 2.2e-16 which is less than 0.05, we reject the null hypothesis and interpret that the regression model is significant.

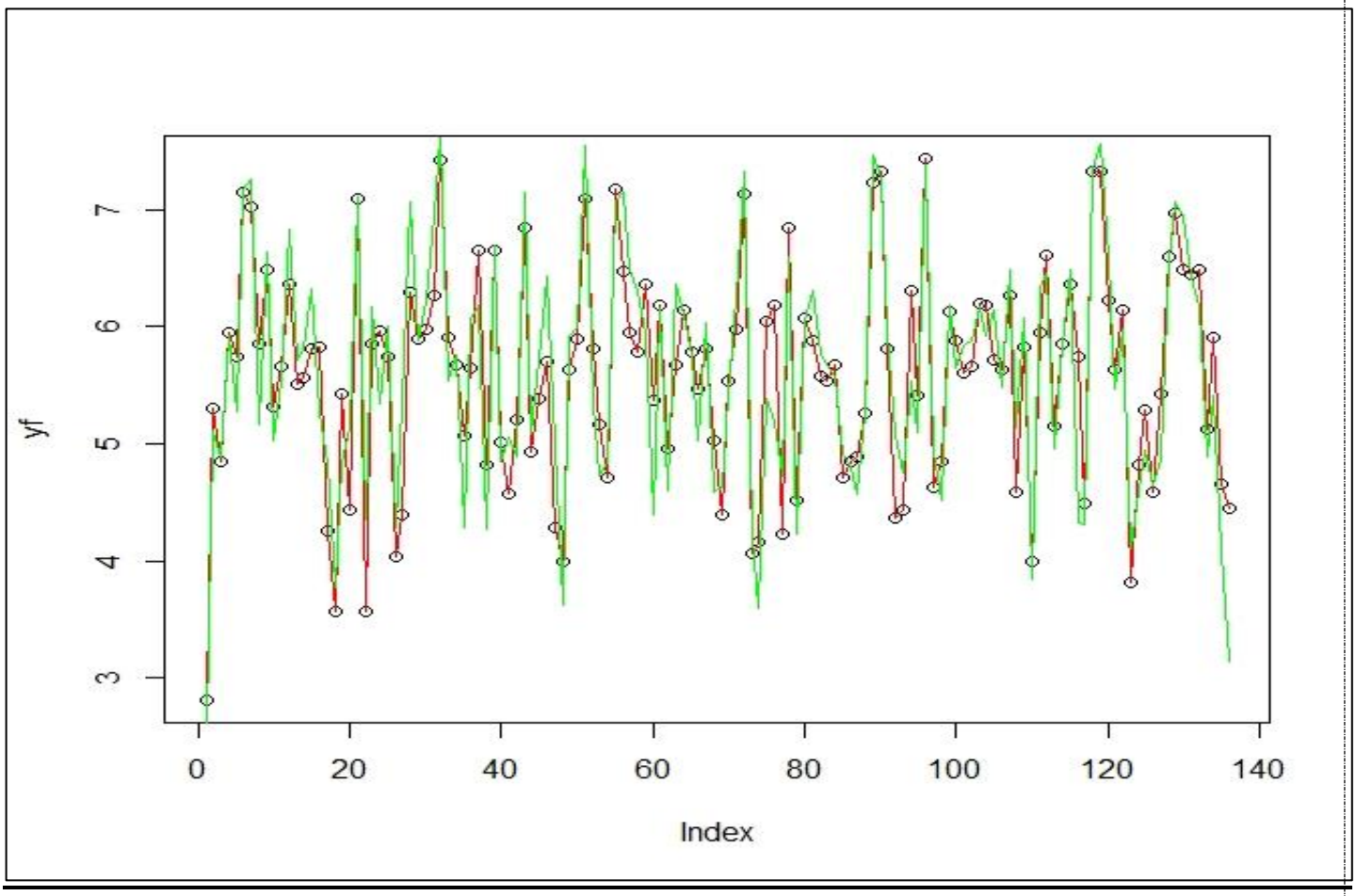
Plot of predicted vs original Ladder Score

Ridge regression model



Plot of predicted vs original Ladder Score

PCR Model



R Codes:

```
rdata <- read.csv("happy.csv")


y<- rdata$`Ladder.score`

x1<-rdata$`Logged.GDP.per.capita`

x2<-rdata$`Social.support`

x3<-rdata$`Healthy.life.expectancy`

x4<-rdata$`Freedom.to.make.life.choices`

x5<-rdata$`Generosity`

x6<-rdata$`Perceptions.of.corruption`


x <- cbind(x1,x2,x3,x4,x5,x6)


frame <- data.frame(y,x1,x2,x3,x4,x5,x6)

print(frame)

pairs(frame)


#checking skewness of independent variables

par(mfrow=c(2,2))
```

```
boxplot(x1)
plot(density(x1))
polygon(density(x1),col= "orange")

boxplot(x2)
plot(density(x2))
polygon(density(x2),col= "orange")

boxplot(x3)
plot(density(x3))
polygon(density(x3),col= "orange")

boxplot(x4)
plot(density(x4))
polygon(density(x4),col= "orange")

boxplot(x5)
plot(density(x5))
polygon(density(x5),col= "orange")
```

```
boxplot(x6)

plot(density(x6))

polygon(density(x6),col= "orange")


library(e1071)

skewness(x1)

skewness(x2)

skewness(x3)

skewness(x4)

skewness(x5)

skewness(x6)


#initial model

model=lm(y~x1+x2+x3+x4+x5+x6)

summary(model)


par(mfrow=c(2,2))

plot(model)

res <- residuals(model)
```

```
#correlation plot

par(mfrow=c(1,1))

library(corrplot)

corrplot(cor(frame),method="number")


#checking and removing outliers

library(olsrr)

ols_plot_cooksd_bar(model)

frame2 <- frame[-c(13,16,41,53,56,63,75,94,115,120,132,137,147),]

print(frame2)

write.csv(frame2,file = "ridge.csv")



ndata <- read.csv("ridge.csv")

y_1<- ndata$`y`

x1_1<-ndata$`x1`

x2_1<-ndata$`x2`

x3_1<-ndata$`x3`

x4_1<-ndata$`x4`

x5_1<-ndata$`x5`
```

```
x6_1<-ndata$`x6`

#model after removal of outliers

newmodel=lm(y_1~x1_1+x2_1+x3_1+x4_1+x5_1+x6_1)

summary(newmodel)

par(mfrow=c(1,1))

plot(newmodel)

res <- residuals(newmodel)

#Checking normality assumption

shapiro.test(res)


#vif of model

library(car)

car::vif(newmodel)


newframe <- data.frame(x1_1,x2_1,x3_1,x4_1,x5_1,x6_1)

#Ridge regression model

library(lmridge)

mod = lmridge(y_1~x1_1+x2_1+x3_1+x4_1+x5_1+x6_1,newframe, K = seq(0,
0.15, 0.002))

par(mfrow=c(1,1))

plot(mod, type = "ridge", abline = FALSE) #stabilizes near 0.03
```

```
kest(mod)$HKB #0.02976505 (Hoerl- kennard iterative method)

model1=lmridge(y_1~x1_1+x2_1+x3_1+x4_1+x5_1+x6_1,newframe,
K=0.02976505)

summary(model1)

library(car)

vif(model1)


#hetroscedasticity of ridge regression model

par(mfrow=c(1,1))

plot(y_1,residuals(model1))

abline(h=0)

plot(x1_1,residuals(model1))

abline(h=0)

plot(x2_1,residuals(model1))

abline(h=0)

plot(x3_1,residuals(model1))

abline(h=0)

plot(x4_1,residuals(model1))

abline(h=0)

plot(x5_1,residuals(model1))
```

```
abline(h=0)

plot(x6_1,residuals(model1))

abline(h=0)


#Glejser's test for checking homoscedasticity

are <- abs(residuals(model1))

summary(lm(are ~ x4_1, data = newframe)) #0.004883 ---highest

summary(lm(are ~ I(1/x4_1), data = newframe)) #0.001402

summary(lm(are ~ sqrt(x4_1), data = newframe))#0.003903

summary(lm(are ~ I(1/sqrt(x4_1)), data = newframe))#0.00213


summary(lm(are ~ x5_1, data = newframe)) #8.104e-08

summary(lm(are ~ I(1/x5_1), data = newframe)) # 0.0008338 ---highest


summary(lm(are ~ x6_1, data = newframe)) #0.04368

summary(lm(are ~ I(1/x6_1), data = newframe)) #0.04796

summary(lm(are ~ sqrt(x6_1), newframe)) #0.04748

summary(lm(are ~ I(1/sqrt(x6_1)), data = newframe))#0.05012 ----highest
```

```
#Goldfeld-Quandt test for checking homoscedasticity

library(lmtest)

gqtest(model1,fraction=1/3,order.by = ~ x4_1,alternative = c("less"))#p-value=0.09882

gqtest(model1,fraction=1/3,order.by = ~ I(1/x5_1),alternative = c("less"))#p-value=0.4269

gqtest(model1,fraction=1/3,order.by = ~ I(1/sqrt(x6_1)),alternative = c("less"))#p value=0.1265


#bruesch pagan test for checking homoscedasticity

bptest(model1)

#R_squared predicted for ridge regression model

tss=sum((y_1-mean(y_1))^2)

r_sq_pred= function(m) {

  1-sum((press.lmridge(m))^2)/tss

}

r_sq_pred(model1) #80.86%
```

```
#PCR Model
rdata <- read.csv("ridge.csv")
xnew=cbind(rdata$x1,rdata$x2,rdata$x3,rdata$x4,rdata$x5,rdata$x6)
pcr=prcomp(xnew,center = T,scale=T)
plot(pcr,type='l')
summary(pcr)

#using pcr model
model2=lm(rdata$y~PC1+PC2+PC3+PC4,data=data.frame(pcr$x))
summary(model2)
car::vif(model2)

PC1=pcr$x[,1]
PC2=pcr$x[,2]
PC3=pcr$x[,3]
PC4=pcr$x[,4]
#heteroscedasticity in PCR model
library(MASS)
par(mfrow=c(1,1))
plot(model2)
```

```
par(mfrow=c(1,1))
plot(PC1,stdres(model2))
abline(h=0)
plot(PC2,stdres(model2))
abline(h=0)
plot(PC3,stdres(model2))
abline(h=0)
plot(PC4,stdres(model2))
abline(h=0)

#bruesch pagan test
bptest(model2)

#PRESS
PRESS <- function(linear.model) {
  #' calculate the predictive residuals
  pr <- residuals(linear.model)/(1-lm.influence(linear.model)$hat)
  #' calculate the PRESS
  PRESS <- sum(pr^2)
```

```
  return(PRESS)
}

#R-sq predicted for PCR Model
pred_r_squared <- function(linear.model) {
  #' Use anova() to get the sum of squares for the linear model
  lm.anova <- anova(linear.model)
  #' Calculate the total sum of squares
  tss <- sum(lm.anova$'Sum Sq')
  # Calculate the predictive R^2
  pred.r.squared <- 1-PRESS(linear.model)/(tss)

  return(pred.r.squared)
}

pred_r_squared(model2)#81.1%
```

#Plotting predicted vs original ladder score

#Ridge Model

yfit=-1.5781 + 0.2390*x1_1 + 2.4325*x2_1 + 0.0298*x3_1 + 2.1707*x4_1 + 0.4991*x5_1 - 0.9676*x6_1

plot(yfit,col='black')

lines(yfit,type='l',col='red')

lines(y_1,type='l',col='green')

#PCR model

yf=5.612941 + 0.509072*PC1 + 0.112711*PC2 - 0.057716*PC3 - 0.001522*PC4

plot(yf,col='black')

lines(yf,type='l',col='red')

lines(y_1,type='l',col='green')

Bibliography

1. Lecture notes of Dr. Sharmishtha Mitra, Associate Professor , Department of Mathematics & Statistics, IIT Kanpur.

2. Introduction To Linear Regression Analysis - Montgomery, Peck, Vining.

3. (Wikipedia) https://en.wikipedia.org/

4. https://stackexchange.com/