MATH 1312 - Regression Analysis

Assignment 2

Anirudhda Pardhi - s3807109

Importing required libraries and reading the data

```
library(car)
```

```
## Loading required package: carData
```

```
library(Hmisc)
```

```
## Loading required package: lattice

## Loading required package: survival

## Loading required package: Formula

## Loading required package: ggplot2

##
## Attaching package: 'Hmisc'

## The following objects are masked from 'package:base':
##
##     format.pval, units
```

```
library(plyr)
```

```
##
## Attaching package: 'plyr'

## The following objects are masked from 'package:Hmisc':
##
##     is.discrete, summarize
```

```
library(tidyr)
library(magrittr)
```

```
##
## Attaching package: 'magrittr'
```

```
## The following object is masked from 'package:tidyr':
##
##     extract
```

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:plyr':
##
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize
```

```
## The following objects are masked from 'package:Hmisc':
##
##     src, summarize
```

```
## The following object is masked from 'package:car':
##
##     recode
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
library(QuantPsyc)
```

```
## Loading required package: boot
```

```
##
## Attaching package: 'boot'
```

```
## The following object is masked from 'package:survival':
##
##     aml
```

```
## The following object is masked from 'package:lattice':
##
##     melanoma
```

```
## The following object is masked from 'package:car':
##
##     logit
```

```
## Loading required package: MASS

##
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':
##
##     select

##
## Attaching package: 'QuantPsyc'

## The following object is masked from 'package:base':
##
##     norm
```

```
library(TSA)
```

```
##
## Attaching package: 'TSA'

## The following objects are masked from 'package:stats':
##
##     acf, arima

## The following object is masked from 'package:utils':
##
##     tar
```

```
data1 <- read.csv("/Users/ADMIN/Desktop/Sem 3/Time Series/Assignment/data1.csv",header=FALSE)
class(data1)
```

```
## [1] "data.frame"
```

Q1

Design matrix

```
liver <- read.csv("/Users/ADMIN/Desktop/Sem 3/Regression analysis/Asg 2/liver.csv")
# Design matrix
abc<- model.matrix(Y ~ BdyWt + LvrWt + Dose , data= liver)
str(abc)
```

```
##  num [1:19, 1:4] 1 1 1 1 1 1 1 1 1 1 ...
##  - attr(*, "dimnames")=List of 2
##   ..$ : chr [1:19] "1" "2" "3" "4" ...
##   ..$ : chr [1:4] "(Intercept)" "BdyWt" "LvrWt" "Dose"
##  - attr(*, "assign")= int [1:4] 0 1 2 3
```

```
abc
```

```
##    (Intercept) BdyWt LvrWt Dose
## 1            1   176   6.5 0.88
## 2            1   176   9.5 0.88
## 3            1   190   9.0 1.00
## 4            1   176   8.9 0.88
## 5            1   200   7.2 1.00
## 6            1   167   8.9 0.83
## 7            1   188   8.0 0.94
## 8            1   195  10.0 0.98
## 9            1   176   8.0 0.88
## 10           1   165   7.9 0.84
## 11           1   158   6.9 0.80
## 12           1   148   7.3 0.74
## 13           1   149   5.2 0.75
## 14           1   163   8.4 0.81
## 15           1   170   7.2 0.85
## 16           1   186   6.8 0.94
## 17           1   146   7.3 0.73
## 18           1   181   9.0 0.90
## 19           1   149   6.4 0.75
## attr(,"assign")
## [1] 0 1 2 3
```

```
X <- cbind(constant = 1, as.matrix(liver[,])[, -4])
Xtrans <- t(X)
Xti <- solve(t(X)%*%X)
Xti
```

```
##              constant        BdyWt        LvrWt       Dose
## constant  6.33809378 -0.074426957 -0.068005387   8.133435
## BdyWt    -0.07442696  0.010644476 -0.002783671  -2.006297
## LvrWt    -0.06800539 -0.002783671  0.049620475   0.183175
## Dose      8.13343533 -2.006296702  0.183175008 388.083181
```

line of best fit

```
lm.fit1<-lm(Y ~ BdyWt + LvrWt + Dose , data= liver)
summary(lm.fit1)
```

```
##
## Call:
## lm(formula = Y ~ BdyWt + LvrWt + Dose, data = liver)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.100557 -0.063233  0.007131  0.045971  0.134691
##
## Coefficients:
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.265922   0.194585   1.367   0.1919
## BdyWt       -0.021246   0.007974  -2.664   0.0177 *
## LvrWt        0.014298   0.017217   0.830   0.4193
## Dose         4.178111   1.522625   2.744   0.0151 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.07729 on 15 degrees of freedom
## Multiple R-squared:  0.3639, Adjusted R-squared:  0.2367
## F-statistic:  2.86 on 3 and 15 DF,  p-value: 0.07197
```

Equation of best fit line is : Y = 0.265 - 0.021 b1 + 0.014 b2 + 4.178 b3 where b1,b2,b3 are slopes for predictors variables BdyWt, LvrWt, Dose.

p-value of equation line is 0.07 which is greater than 0.05 and this suggest that regression is statistically insignificant at 5% level of significance.

T-test :

Since the sample size is < 30 we need to use T-distribution for 95% confidence. For Degree of freedom = 12 the obtained critical value from T-table is Tc = 2.13. As per the description given in summary we can observe that only T-value for BdyWt= 2.664 and Dose=2.744 is greater than critical T-value hence we can say that BdyWt,Dose are the significant predictor variables. Similarly LvrWt t-value= 0.83 is lower than critical T-value hence we can say that it is insignificant.

Anova table

```
anova(lm.fit1)
```

```
## Analysis of Variance Table
##
## Response: Y
##           Df   Sum Sq  Mean Sq F value  Pr(>F)
## BdyWt      1 0.003216 0.003216  0.5383 0.47446
## LvrWt      1 0.003067 0.003067  0.5134 0.48467
## Dose       1 0.044982 0.044982  7.5296 0.01507 *
## Residuals 15 0.089609 0.005974
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As per ANOVA table p-value of slope for variable BdyWt = 0.47 and LvrWt = 0.48 are greater than 0.05 which suggest that they are statistically not significant at 5% level of significance. However slope of p-value for Dose = 0.015 is less than 0.05 which suggest that this is statistically significant. Similar conclusion can be made with the help of F-value as in case of BdyWt and LvrWt F-value is very low and for slope of Dose the F-value is higher than critical F-value.

```
lm.fit2<-lm(Y ~ BdyWt + Dose , data= liver)
summary(lm.fit2)
```

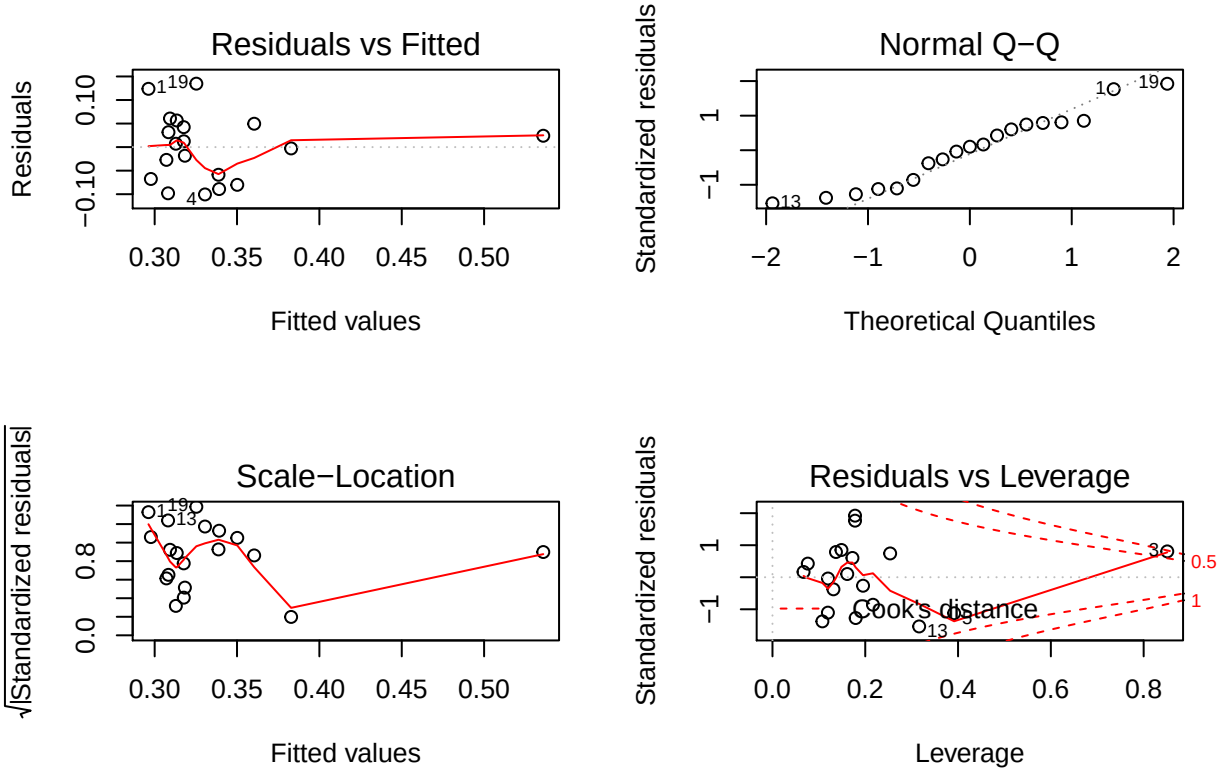
```
##
## Call:
```

```
## lm(formula = Y ~ BdyWt + Dose, data = liver)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.12333 -0.07416  0.01238  0.04884  0.12668
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.285517   0.191267   1.493   0.1550
## BdyWt       -0.020444   0.007838  -2.608   0.0190 *
## Dose         4.125329   1.506472   2.738   0.0146 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.07654 on 16 degrees of freedom
## Multiple R-squared:  0.3347, Adjusted R-squared:  0.2515
## F-statistic: 4.024 on 2 and 16 DF,  p-value: 0.0384
```

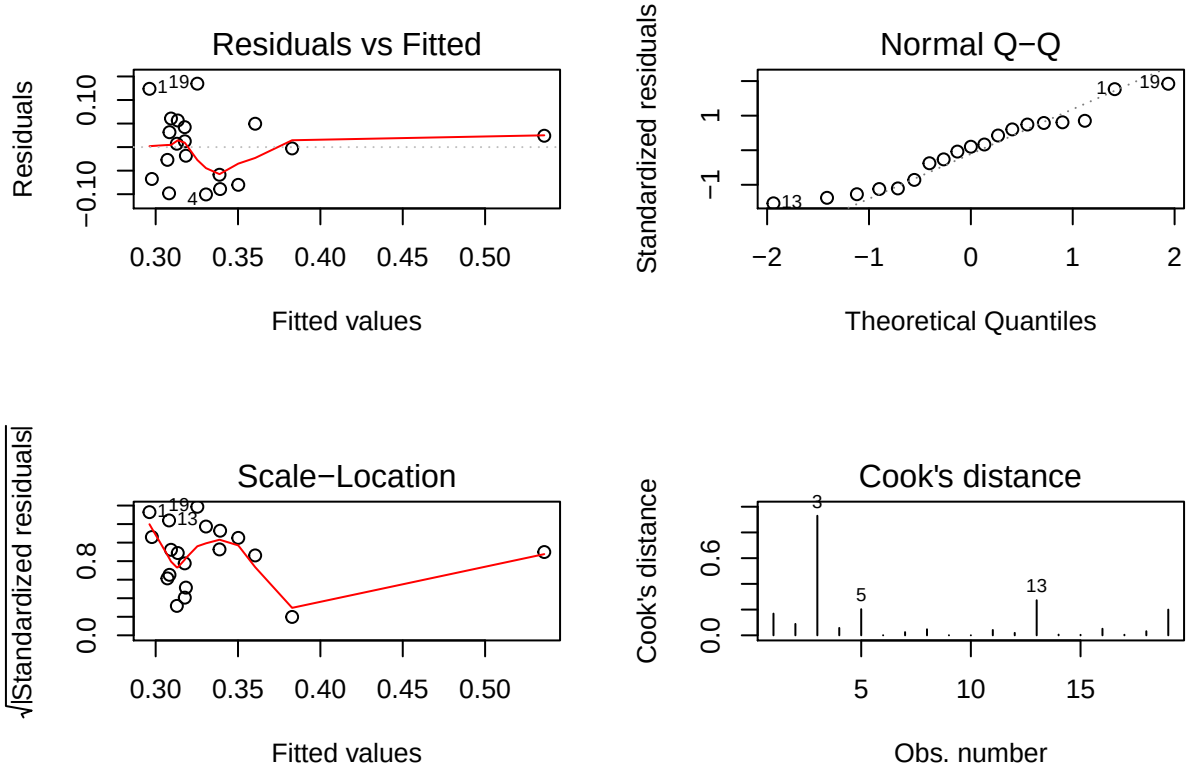
p-value of equation line is 0.03 which is lower than 0.05 and it suggest that regression is statistically significant.Adjusted R-squared value also got increased when variable LvrWt was removed.

Residual analysis

```
par(mfrow=c(2,2))
plot(lm.fit1)
```




```
plot(lm.fit1, which = 1:4)
```




```
ncvTest(lm.fit1)
```

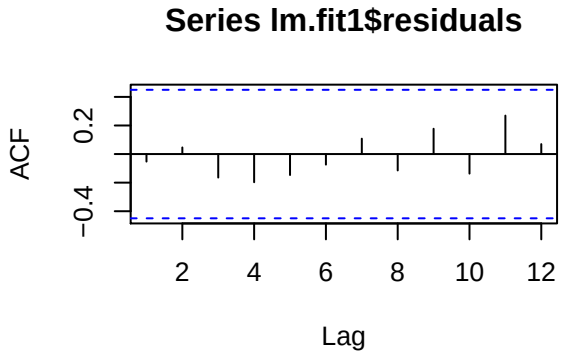
```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 0.6274991, Df = 1, p = 0.42827
```

```
shapiro.test(lm.fit1$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  lm.fit1$residuals
## W = 0.9515, p-value = 0.4189
```

```
acf(lm.fit1$residuals)
durbinWatsonTest(lm.fit1)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1     -0.05276333      1.732152   0.614
##  Alternative hypothesis: rho != 0
```



NCV test- In residual vs fitted graph we can see that the red line is curved so there may be heteroscedasticity exists.So we do "NCV test" and p = 0.428 which is more than significance level 0.05 we can reject the null hypothesis that the variance of the residuals is constant and can say that Homoscedasticity is present.

Shapiro test- To test the Normality we can see the QQ plot and can say that there is not any gross deviations from normality.But since the number of observations are less than 30 it is safe to do "Shapiro test".Obtained p value = 0.41 which is greater than significance level 0.05 which is implying that the distribution of the data are not significantly different from normal distribution.Thus we can assume the normality.

ACF test- In ACF we check for early lags.Before lag = 5 we can observe that no correlation values are crossing significant confidence boundaries hence we can comprehend that stochastic component of data is white noise.As per durbinWatsonTest result we can see that p value is 0.62 which is greater than 0.05 and suggests we fail to reject null hypothesis i.e First-order autocorrelation does not exist.

Cook's distance shows the presence of influential points or possible outliers.In our case we have spotted one such point at 3rd obsrvation.

check for multicollinearity

```
predictor <- liver[,-c(4)]
cor(predictor)
```

```
##           BdyWt     LvrWt      Dose
## BdyWt 1.0000000 0.5000101 0.9902126
## LvrWt 0.5000101 1.0000000 0.4900711
## Dose  0.9902126 0.4900711 1.0000000
```

```r
lm.fit1<-lm(Y ~ BdyWt + LvrWt + Dose , data= liver)
lm.fit2<-lm(Y ~ BdyWt + Dose , data= liver)
anova(lm.fit2,lm.fit1)
```

```
## Analysis of Variance Table
##
## Model 1: Y ~ BdyWt + Dose
## Model 2: Y ~ BdyWt + LvrWt + Dose
##   Res.Df      RSS Df Sum of Sq      F Pr(>F)
## 1     16 0.093729
## 2     15 0.089609  1   0.00412 0.6897 0.4193
```

We are checking multicollinearity between the predictors using the correlation matrix.As per this correlation matrix the linear correlation between pair of variables BdyWt and Dose is 99.02% which is very high.

According to ANOVA table we can see that Residual sum of squares for predictorvariable LivWt accounts for only 0.004 and p-value = 0.41 suggests that we cannot reject null hypothesis which indicates slope(beta) for LvrWt = 0.

Backward Elimination

```r
step(lm.fit1, data=liver, direction="backward")
```

```
## Start:  AIC=-93.78
## Y ~ BdyWt + LvrWt + Dose
##
##         Df Sum of Sq      RSS     AIC
## - LvrWt  1  0.004120 0.093729 -94.924
## <none>               0.089609 -93.778
## - BdyWt  1  0.042408 0.132017 -88.416
## - Dose   1  0.044982 0.134591 -88.049
##
## Step:  AIC=-94.92
## Y ~ BdyWt + Dose
##
##         Df Sum of Sq      RSS     AIC
## <none>               0.093729 -94.924
## - BdyWt  1  0.039851 0.133580 -90.192
## - Dose   1  0.043929 0.137658 -89.621

##
## Call:
## lm(formula = Y ~ BdyWt + Dose, data = liver)
##
## Coefficients:
## (Intercept)        BdyWt         Dose
##     0.28552     -0.02044      4.12533
```

In order to find smaller parsimonious model we are applying Backward Elimination method. In this method we prefer to select predictor variables with higher AIC values.

As per the result we conclude that in the multiple linear regression model for a response variable there are 2 significant predictor variable which are BdyWt, Dose and both have high collinearity.

Q2

```
crest <- read.csv("/Users/ADMIN/Desktop/Sem 3/Regression analysis/Asg 2/crest.csv")
mcrest <- crest[,-c(1)]
# line of best fit
crest.fit1<-lm(Y ~ X1 + X2 + X3 , data= crest)
summary(crest.fit1)
```

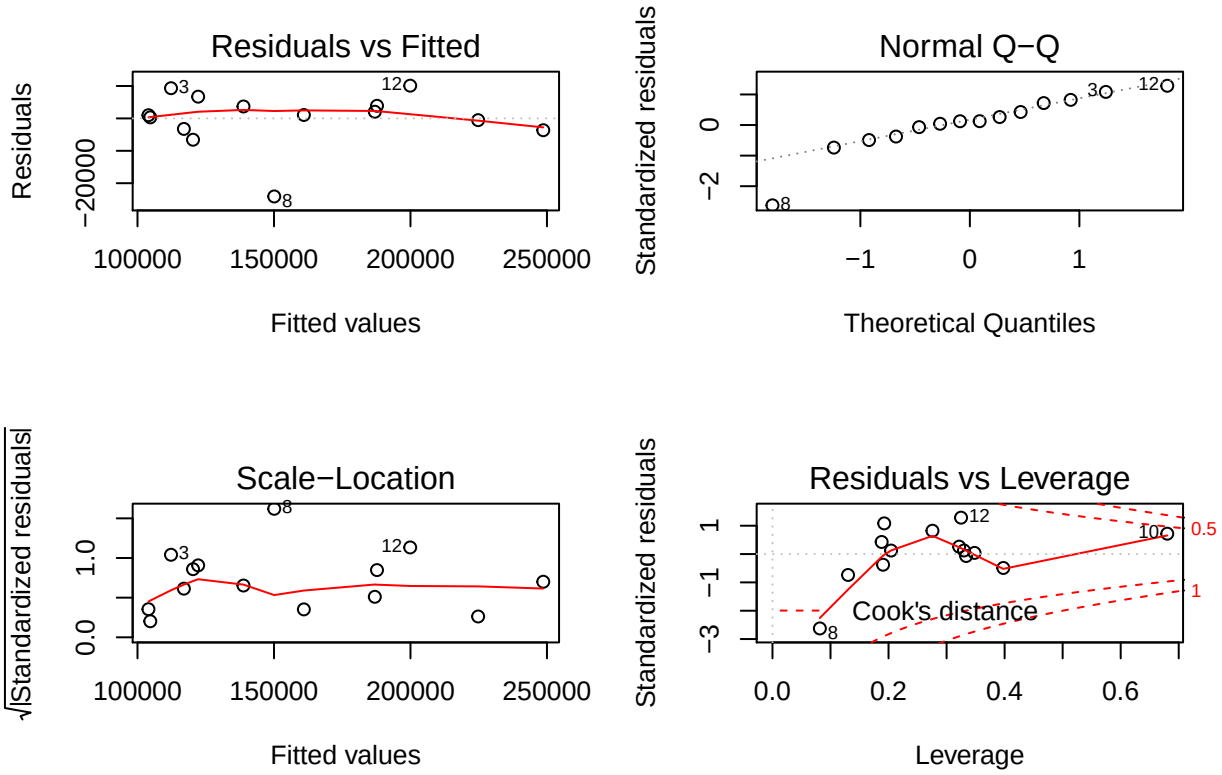
```
##
## Call:
## lm(formula = Y ~ X1 + X2 + X3, data = crest)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -24088  -2568   1021   3836  10100
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  34104.559  17654.144   1.932 0.082187 .
## X1               3.746      1.976   1.896 0.087243 .
## X2          -30046.343  22859.674  -1.314 0.218066
## X3              85.926     17.911   4.797 0.000727 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9574 on 10 degrees of freedom
## Multiple R-squared:  0.9691, Adjusted R-squared:  0.9598
## F-statistic: 104.5 on 3 and 10 DF,  p-value: 7.537e-08
```

Equation of best fit line is : Y = 34104.559 +3.746 X1 - 30046.343 X2 + 85.926 X3 where X1,X2,X3 are slopes for variables.

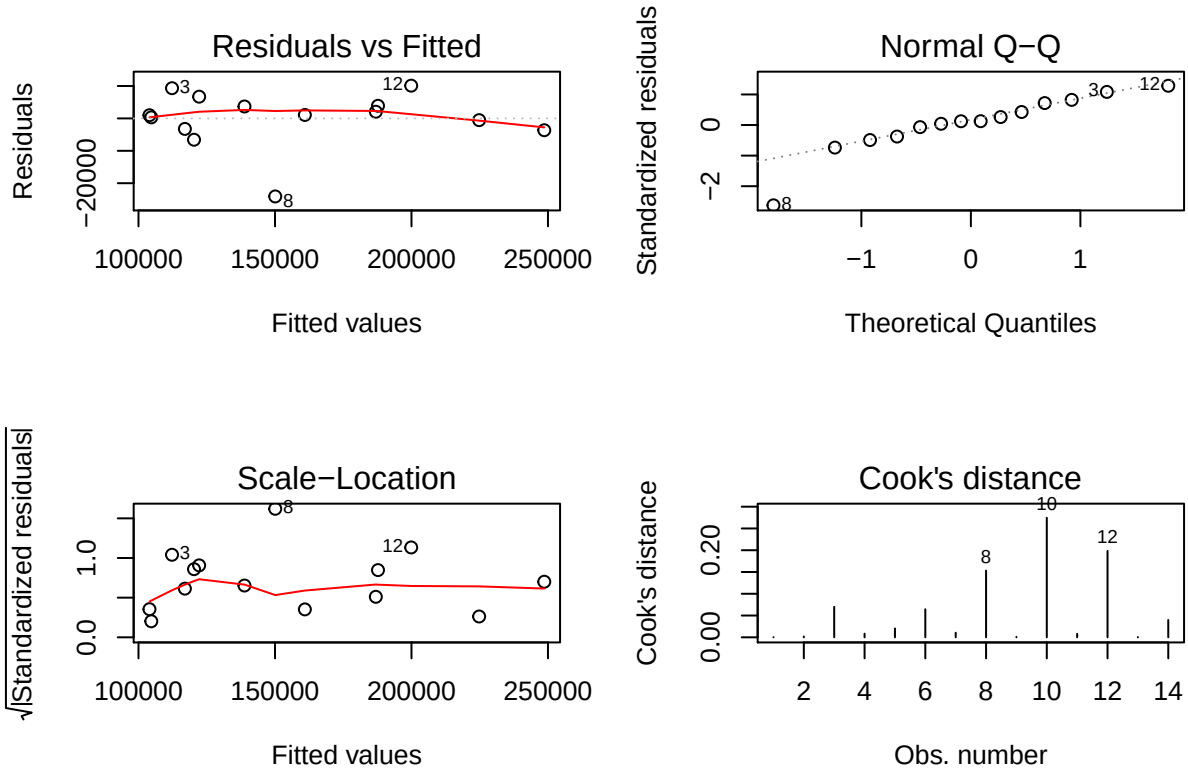
p-value of equation line is 7.537e-08 which is less than 0.05 and it suggest that regression is significant.

Residual analysis

```
par(mfrow=c(2,2))
plot(crest.fit1)
```




```
plot(crest.fit1, which = 1:4)
```




```
ncvTest(crest.fit1)
```

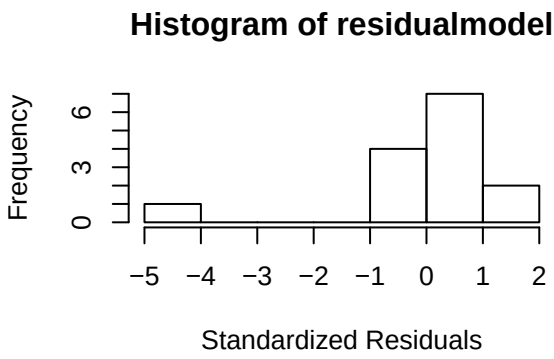
```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 0.07424735, Df = 1, p = 0.78525
```

```
shapiro.test(crest.fit1$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  crest.fit1$residuals
## W = 0.83777, p-value = 0.01522
```

```
residualmodel = rstudent(crest.fit1)
hist(residualmodel,xlab="Standardized Residuals")
acf(crest.fit1$residuals)
durbinWatsonTest(crest.fit1)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1      -0.1129573      2.211283   0.882
##  Alternative hypothesis: rho != 0
```



NCV test- In residual vs fitted graph we can see that the red line is curved and the residuals seem to increase as the fitted Y values increase so there may be heteroscedasticity exists.So we do "NCV test" and p-value = 0.78 which is greater than significance level 0.05 we can reject the null hypothesis that the variance of the residuals is constant and can say that Homoscedasticity is present.

Shapiro test- To test the Normality we can see the QQ plot and can say that initially there was a point deviating from normality.But since the number of observations are less than 30 it is safe to do "Shapiro test".Obtained p value = 0.015 which is less than 0.05 which is implying that the distribution of the data are different from normal distribution.If we observe the histogram of residual distribution it does not seems to be normal thus we cannot assume the normality and this assumption is violated.

ACF test- In ACF we check for early lags.Before lag = 5 we can observe that no correlation values are crossing significant confidence boundaries hence we can comprehend that stochastic component of data is white noise.As per durbinWatsonTest result we can see that p value is 0.88 which is greater than 0.05 and suggests we fail to reject null hypothesis i.e First-order autocorrelation does not exist.

Cook's distance shows the presence of influential points or possible outliers.In our case we have spotted one such point at 10th obsrvation.

check for multicollinearity

```
crestpredictor <- crest[,-c(1)]
cor(crestpredictor)
```

```
##            Y        X1        X2        X3
## Y  1.0000000 0.9292781 0.5988912 0.9787569
```

```
## X1 0.9292781 1.0000000 0.7714611 0.9187616
## X2 0.5988912 0.7714611 1.0000000 0.6154342
## X3 0.9787569 0.9187616 0.6154342 1.0000000
```

We are checking multicollinearity between the predictors using the correlation matrix.As per this correlation matrix the linear correlation between pair of variables X1 and X3 is 91.87% which is very high.

```
anova(crest.fit1)
```

```
## Analysis of Variance Table
##
## Response: Y
##           Df     Sum Sq    Mean Sq F value    Pr(>F)
## X1         1 2.5611e+10 2.5611e+10 279.392  1.23e-08 ***
## X2         1 1.0202e+09 1.0202e+09  11.130 0.0075401 **
## X3         1 2.1097e+09 2.1097e+09  23.015 0.0007265 ***
## Residuals 10 9.1667e+08 9.1667e+07
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

According to ANOVA table p-value of X1,X2,X3 are below 0.05 which indicates all the predictor variables are significant.

Different predictor variable combinations

```
crest.fit1<-lm(Y ~ X1 + X2 + X3 , data= crest)
summary(crest.fit1)
```

```
##
## Call:
## lm(formula = Y ~ X1 + X2 + X3, data = crest)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -24088  -2568   1021   3836  10100
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  34104.559  17654.144   1.932 0.082187 .
## X1               3.746      1.976   1.896 0.087243 .
## X2          -30046.343  22859.674  -1.314 0.218066
## X3              85.926     17.911   4.797 0.000727 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9574 on 10 degrees of freedom
## Multiple R-squared:  0.9691, Adjusted R-squared:  0.9598
## F-statistic: 104.5 on 3 and 10 DF,  p-value: 7.537e-08
```

```r
crest.fit2<-lm(Y ~ X1 + X2 , data= crest)
summary(crest.fit2)
```

```
##
## Call:
## lm(formula = Y ~ X1 + X2, data = crest)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -22435  -7716  -2156  12546  22346
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  16832.108  29941.811   0.562   0.5853
## X1              12.138      1.592   7.625 1.03e-05 ***
## X2          -70800.337  36766.675  -1.926   0.0804 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16590 on 11 degrees of freedom
## Multiple R-squared:  0.898,  Adjusted R-squared:  0.8794
## F-statistic:  48.4 on 2 and 11 DF,  p-value: 3.534e-06
```

```r
crest.fit4<-lm(Y ~ X2 + X3 , data= crest)
summary(crest.fit4)
```

```
##
## Call:
## lm(formula = Y ~ X2 + X3, data = crest)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -25401.1  -1468.2   -170.3   4058.9  17343.5
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 37335.393  19533.674   1.911   0.0824 .
## X2          -1356.392  19045.157  -0.071   0.9445
## X3            115.986      9.259  12.526 7.47e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10640 on 11 degrees of freedom
## Multiple R-squared:  0.958,  Adjusted R-squared:  0.9503
## F-statistic: 125.4 on 2 and 11 DF,  p-value: 2.684e-08
```

As per obtained value from above we can check R-squared value in different cases of variable combination :

Y ~ X1 + X2 + X3 , R2 value = 95.98 % Y ~ X1 + X2 , R2 value = 87.94 % Y ~ X2 + X3 , R2 value = 95.03 %

Since we have already detected the problem of high collinearity between X1 and X3 we can check which variable is can be insignificant among the predictor variable pair. Based on the above R-square value we

can say that when pair of variable were X2 and X3 the model performance was better (95.03%) and since X1 was highly correlated with X3 the model was performing well too when pair of variables were X1 + X2 + X3 (95.98%). When X1 was treated as pair with X2 the model performance (87.94) got dropped. So we can conclude that predictor variable X2 and X3 are significant pair of variables and because of detected multicollinearity problem we can say X1 is less significant in model.

Forward, backward and stepwise regression models

```
mcrest <- crest[,-c(1)]
# Full model shoul contains all the variables
full=lm(Y~., data=mcrest)

# null model contains no variable
null=lm(Y~1, data=mcrest)

#Backward elimination using AIC values
step(full, data=mcrest, direction="backward")
```

```
## Start:  AIC=259.96
## Y ~ X1 + X2 + X3
##
##        Df  Sum of Sq        RSS    AIC
## <none>                916674966 259.96
## - X2    1  158364757 1075039723 260.19
## - X1    1  329414202 1246089168 262.26
## - X3    1 2109684588 3026359554 274.68

##
## Call:
## lm(formula = Y ~ X1 + X2 + X3, data = mcrest)
##
## Coefficients:
## (Intercept)           X1           X2           X3
##   34104.559        3.746   -30046.343       85.926
```

```
#forward selection using AIC values
step(null, scope=list(lower=null, upper=full), direction="forward")
```

```
## Start:  AIC=302.64
## Y ~ 1
##
##        Df  Sum of Sq        RSS    AIC
## + X3    1 2.8411e+10 1.2467e+09 260.27
## + X1    1 2.5611e+10 4.0466e+09 276.75
## + X2    1 1.0637e+10 1.9020e+10 298.42
## <none>               2.9658e+10 302.63
##
## Step:  AIC=260.27
## Y ~ X3
##
##        Df Sum of Sq        RSS    AIC
```

```
## + X1    1 171624035 1075039723 260.19
## <none>              1246663758 260.27
## + X2    1    574590 1246089168 262.26
##
## Step:  AIC=260.19
## Y ~ X3 + X1
##
##        Df Sum of Sq        RSS    AIC
## + X2    1 158364757  916674966 259.96
## <none>              1075039723 260.19
##
## Step:  AIC=259.96
## Y ~ X3 + X1 + X2
```

```
##
## Call:
## lm(formula = Y ~ X3 + X1 + X2, data = mcrest)
##
## Coefficients:
## (Intercept)           X3           X1           X2
##   34104.559       85.926        3.746   -30046.343
```

```
#stepwise regression using AIC values
step(null, scope = list(upper=full), data=mcrest, direction="both")
```

```
## Start:  AIC=302.64
## Y ~ 1
##
##        Df  Sum of Sq        RSS    AIC
## + X3    1 2.8411e+10 1.2467e+09 260.27
## + X1    1 2.5611e+10 4.0466e+09 276.75
## + X2    1 1.0637e+10 1.9020e+10 298.42
## <none>               2.9658e+10 302.63
##
## Step:  AIC=260.27
## Y ~ X3
##
##        Df  Sum of Sq        RSS    AIC
## + X1    1 1.7162e+08 1.0750e+09 260.19
## <none>               1.2467e+09 260.27
## + X2    1 5.7459e+05 1.2461e+09 262.26
## - X3    1 2.8411e+10 2.9658e+10 302.63
##
## Step:  AIC=260.19
## Y ~ X3 + X1
##
##        Df  Sum of Sq        RSS    AIC
## + X2    1  158364757  916674966 259.96
## <none>               1075039723 260.19
## - X1    1  171624035 1246663758 260.27
## - X3    1 2971530267 4046569990 276.75
##
## Step:  AIC=259.96
```

```
## Y ~ X3 + X1 + X2
##
##        Df  Sum of Sq        RSS    AIC
## <none>                916674966 259.96
## - X2    1  158364757 1075039723 260.19
## - X1    1  329414202 1246089168 262.26
## - X3    1 2109684588 3026359554 274.68

##
## Call:
## lm(formula = Y ~ X3 + X1 + X2, data = mcrest)
##
## Coefficients:
## (Intercept)           X3           X1           X2
##   34104.559       85.926        3.746   -30046.343
```

Based on AIC values we have calculated significant predictor variables in regression model using forward, backward and stepwise regression procedures.

Results for different model building procedures are as below :

Backward selection (or backward elimination), which starts with all predictors in the model (full model), iteratively removes the least contributing predictors, and stops when you have a model where all predictors are statistically significant.

Backward elimination : Y ~ X3 + X1 + X2

Forward selection, which starts with no predictors in the model, iteratively adds the most contributive predictors, and stops when the improvement is no longer statistically significant.

Forward selection : Y ~ X3 + X1 + X2

Stepwise selection (or sequential replacement), which is a combination of forward and backward selections. You start with no predictors, then sequentially add the most contributive predictors (like forward selection). After adding each new variable, remove any variables that no longer provide an improvement in the model fit.

Stepwise regression : Y ~ X3 + X1 + X2

As per the above results pair of predictor variable X1, X2, X3 seems to be significant in regression model but as discussed earlier because X1 and X3 have high collinearity the problem of multicollinearity exist in the regression model.