CLATHERATE FORMATION DATA LAB 5

A.Mayuri(2348133)
[COMPANY NAME] [Company address]

Clatherate Formation Data

Lab 5

A.Mayuri(2348133)

2023-12-08

Problem Statement:

1. Fit a suitable linear regression model.
2. Construct a normal probability plot of the residuals. Does there seem to be any problem with the normality and constant variance assumption?If yes, what remedial measure will u perform?
3. Construct and interpret a plot of the residuals.
4. Are the residuals correlated?
5. Is multicollinearity a potential problem in your model? If it is a problem, what is your remedy?
6. Are there any outliers in the data? If it exists, how will you treat it?

Import Dataset

```
library(readxl)
Chem <- read_excel("C:/Users/mayur/Desktop/Mstat/Semesters/Tri-sem2/Regressio
n/Dataset/Chem.xlsx")
View(Chem)
attach(Chem)
```

1) Understanding the Variables Using correlation

```
cor(Chem)

##             x1         x2         y
## x1  1.0000000 -0.1275387 0.5192537
## x2 -0.1275387  1.0000000 0.6838246
## y   0.5192537  0.6838246 1.0000000
```

We observe that there is a positive linear relation between the independent and dependent variable ie, (X1,Y) AND (X2,Y). Also there is a very low correlation between (X1,X2). Hence they are independent to each other.

2) Fitting a Linear Regression Model

```
model1=lm(Chem$y~.,data = Chem)
model1

##
## Call:
## lm(formula = Chem$y ~ ., data = Chem)
##
## Coefficients:
```

```
## (Intercept)           x1           x2
##     11.0870     350.1192       0.1089
```

```
summary(model1)
```

```
##
## Call:
## lm(formula = Chem$y ~ ., data = Chem)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.7716 -4.1656  0.0802  3.8323  8.3349
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.109e+01  1.669e+00   6.642 1.48e-07 ***
## x1          3.501e+02  3.968e+01   8.823 3.38e-10 ***
## x2          1.089e-01  9.983e-03  10.912 1.74e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.782 on 33 degrees of freedom
## Multiple R-squared:  0.8415, Adjusted R-squared:  0.8319
## F-statistic:  87.6 on 2 and 33 DF,  p-value: 6.316e-14
```

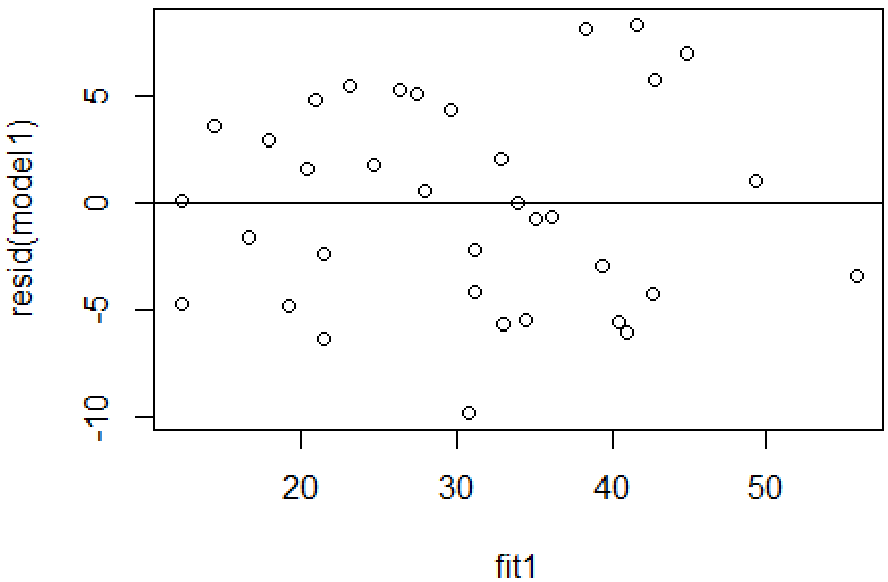
Interpretation: We observe that both the independent variables have a significant linear relationship. since the values of both the independent variables are <=0.05 we reject null hypothesis and accept alternative hypothesis.ie, there exisist a linear relationship between parameter and dependent variable.

3) Residual Analysis (Question3)

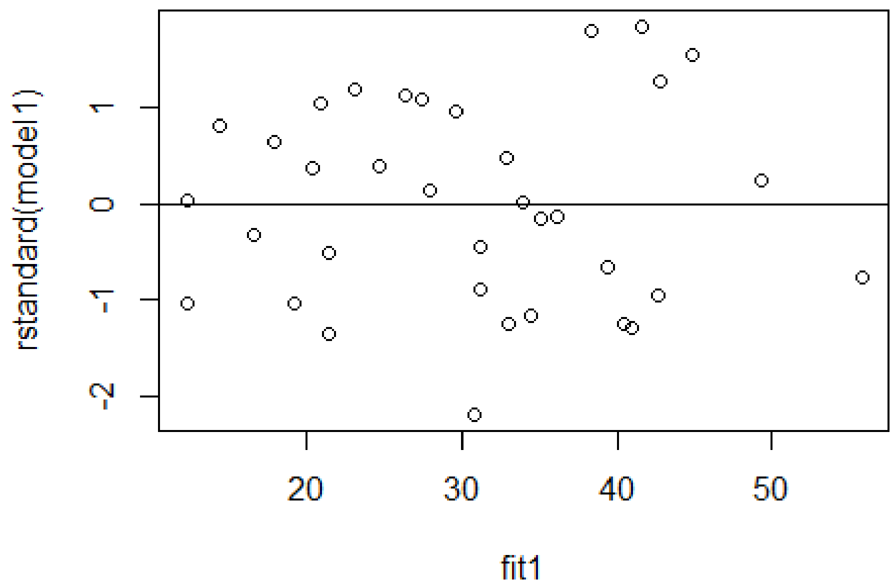
```
fit1=fitted.values(model1)
fit1
```

```
##        1        2        3        4        5        6        7        8
## 12.17632 16.53370 20.34641 23.06977 26.33780 29.60583 32.87386 36.14190
##        9       10       11       12       13       14       15       16
## 39.40993 42.67796 12.17632 14.35501 17.84091 20.89108 27.42714 33.96321
##       17       18       19       20       21       22       23       24
## 40.49927 19.17871 21.35740 24.62543 27.89346 31.16150 40.96559 21.35740
##       25       26       27       28       29       30       31       32
## 24.62543 31.16150 34.42953 30.77163 32.95032 42.75442 49.29048 55.82655
##       33       34       35       36
## 35.12901 38.39704 41.66507 44.93311
```

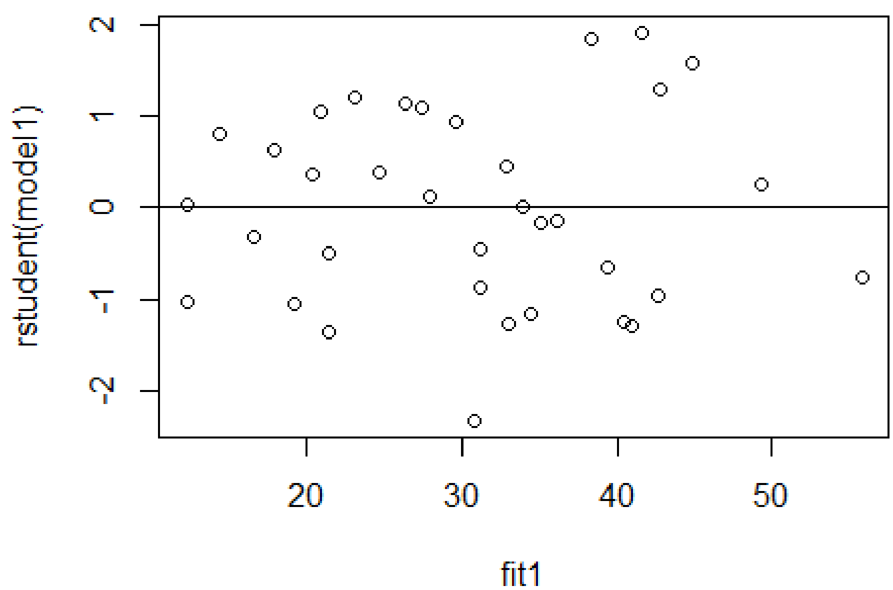
```r
# Residual analysis
plot(fit1,resid(model1))
abline(0,0)
```




```r
plot(fit1,rstandard(model1))#standardized residual model
abline(0,0)
```




```
plot(fit1,rstudent(model1))# studentized residual model
abline(0,0)
```



```
## Residual values
residual=resid(model1)
residual
```

```
##           1           2           3           4           5           6 
## -4.67632456 -1.53370131  1.65359404  5.53023358  5.26220102  4.39416846 
##           7           8           9          10          11          12 
##  2.12613590 -0.64189666 -2.90992922 -4.17796177  0.12367544  3.64498707 
##          13          14          15          16          17          18 
##  2.95908567  4.80892195  5.07285683  0.03679172 -5.49927340 -4.77870947 
##          19          20          21          22          23          24 
## -2.35739785  1.77456959  0.60653704 -2.16149552 -5.96559320 -6.25739785 
##          25          26          27          28          29          30 
##  1.77456959 -4.16149552 -5.42952808 -9.77163103 -5.65031940  5.74558292 
##          31          32          33          34          35          36 
##  1.10951780 -3.32654731 -0.72900778  8.10295966  8.33492711  6.96689455
```

Interpretation: Using standard and studentized residual plot we observe that there is no pattern hence we cannot comment about both the assumption . we will further check it by using the test.

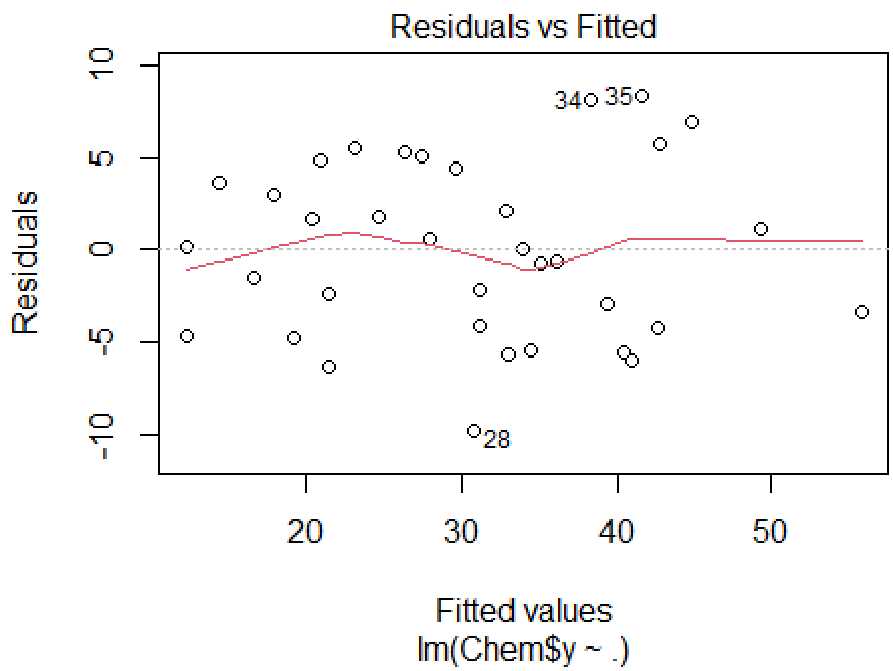
4a)Normality (Question2) and Variance

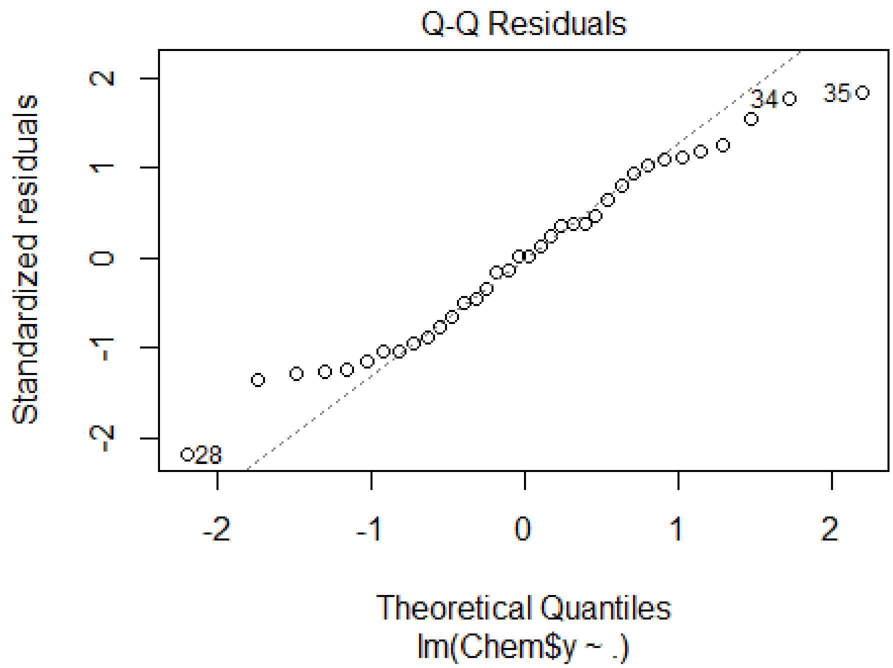
```
# to check for res vales (Normality check)
re1=rstandard(model1)
re1
```

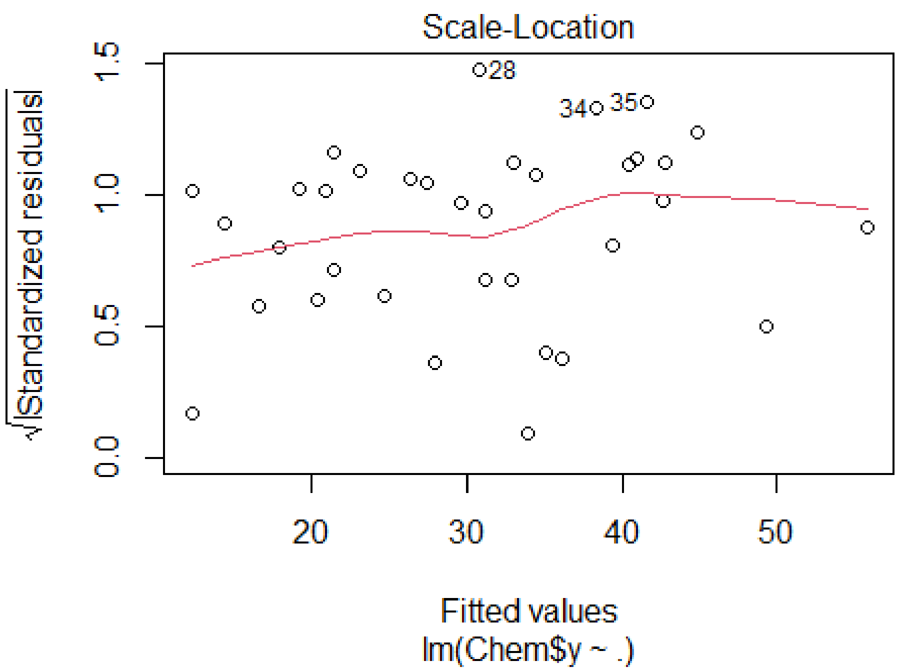
```
##            1            2            3            4            5 
  6 
## -1.037115175 -0.333758043  0.356275584  1.187266673  1.129176925  0.946369
325 
##            7            8            9           10           11 
 12 
##  0.461534357 -0.141070312 -0.650555446 -0.955185630  0.027428737  0.799919
557 
##           13           14           15           16           17 
 18 
##  0.641316052  1.035126037  1.089371414  0.008015465 -1.237884530 -1.041331
230 
##           19           20           21           22           23 
 24 
## -0.508984388  0.379286883  0.128876651 -0.458446375 -1.289781318 -1.351031
103 
##           25           26           27           28           29 
 30 
##  0.379286883 -0.882640059 -1.154169443 -2.186886007 -1.255401518  1.266420
897
```

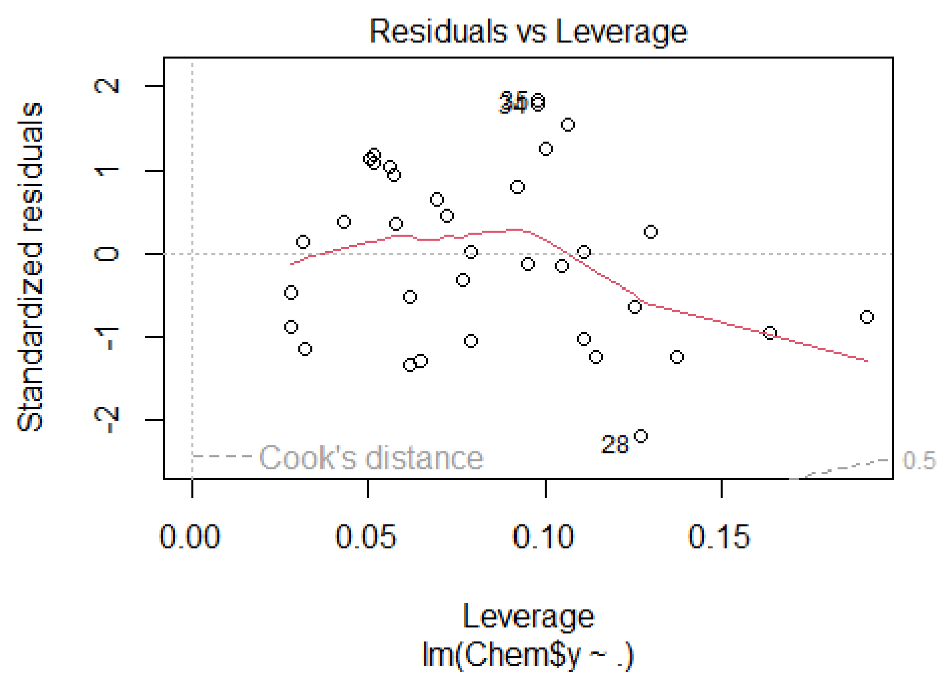
```
##           31           32           33           34           35
 36
##  0.248696768 -0.773278110 -0.161131924  1.783666308  1.835215764  1.541125
067
```

```r
plot(model1)
```


Residuals vs Fitted
Residuals
Fitted values
lm(Chem$y ~ .)
34
35
28
10
5
0
-5
-10
20
30
40
50


Q-Q Residuals
Standardized residuals
Theoretical Quantiles
lm(Chem$y ~ .)
34
35
28
2
1
0
-1
-2






```
shapiro.test(re1)

##
##  Shapiro-Wilk normality test
```

```
##
## data:  re1
## W = 0.97171, p-value = 0.474
```

Interpretation: For Normality assumption using Shapiro Wilk Test at 0.05 level of significance the p value>=0.05 thus we fail to reject null thus the residual follow normal distribution hence the assumption of errors. This is also confirmed using the QQ residual plot

4b) Hypothesis testing for constant variance using BP test :

test for errors have constant variance through errors. use the fitted model.

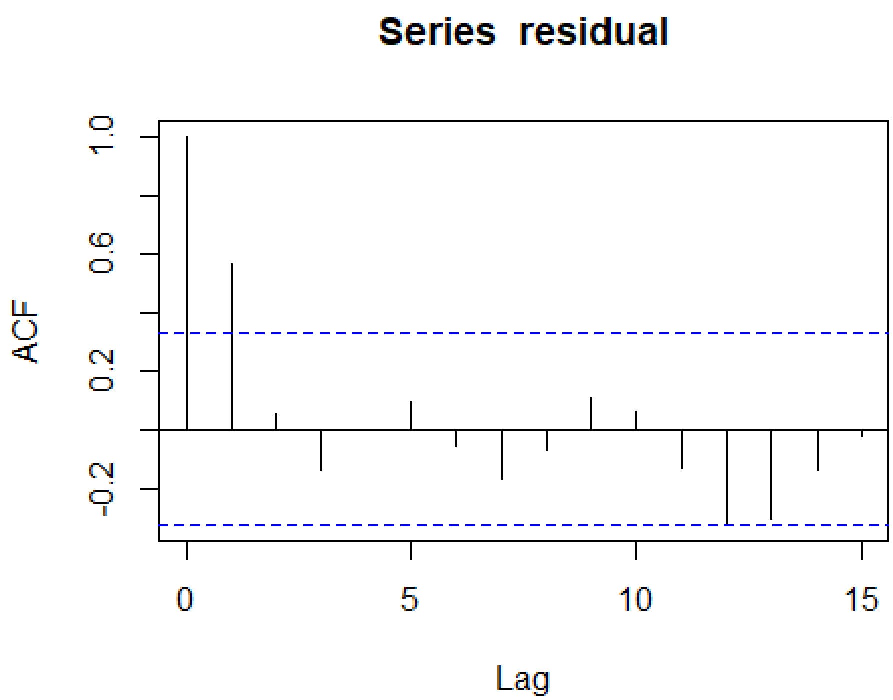
h0: error have const variance h1: error have not const variance

```
library(lmtest)

## Loading required package: zoo

##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric

bptest(model1)

##
##  studentized Breusch-Pagan test
##
## data:  model1
## BP = 8.0945, df = 2, p-value = 0.01747
```

Interpretation: since p-value is <=0.05 we reject h0 ,thus there is no constant variance . hence the assumption of constant variance is not validated . we can perform a log transformation to the dependent variable and try to redefine the model.

5a) Autocorrelation (question4)

```
acf(residual)
```



Interpretation: ACF at 0 is always 1. and all acf points are not within the threshold lines from lag 1 it indicates that there is a significant autocorrelation among the residual series. However we can also confirm the same with durbin watson test procdure.

b) Durbin watson for ACF

h0: rho=0 there is no autocorrelation h1: rho=!0 there is autocorrelation

```
dwtest(model1)
```

```
##
##  Durbin-Watson test
##
## data:  model1
## DW = 0.77943, p-value = 6.004e-06
## alternative hypothesis: true autocorrelation is greater than 0
```

Interpretation: at 5 % level of significance, the p value (6.004e-06)<0.05, we reject the null hypothesis that there is a significant auto-correlation. ie,rho=!=0.

6)Multi-collilinearity (Question5):

We observe that the there could be no multi-collilinearity between independent variable since the correlation between them is (-0.12)<0.7. hence their VIF would be less than 5.

For confirmation,

```
library(car)

## Loading required package: carData

vif(model1)

##       x1       x2
## 1.016535 1.016535
```

Interpretation: As stated the VIF<=5 thus there is no multi-collilinearity.

7)Outliers (Question6)

```
rstandard(model1)

##            1            2            3            4            5 
  6 
## -1.037115175 -0.333758043  0.356275584  1.187266673  1.129176925  0.946369
325 
##            7            8            9           10           11 
 12 
##  0.461534357 -0.141070312 -0.650555446 -0.955185630  0.027428737  0.799919
557 
##           13           14           15           16           17 
 18 
##  0.641316052  1.035126037  1.089371414  0.008015465 -1.237884530 -1.041331
230 
##           19           20           21           22           23 
 24 
## -0.508984388  0.379286883  0.128876651 -0.458446375 -1.289781318 -1.351031
103 
##           25           26           27           28           29 
 30 
##  0.379286883 -0.882640059 -1.154169443 -2.186886007 -1.255401518  1.266420
897 
##           31           32           33           34           35 
 36 
##  0.248696768 -0.773278110 -0.161131924  1.783666308  1.835215764  1.541125
067
```

Interpretation: Here we observe that there is no observation below -3 and above 3. hence there are no outliers.