# Linear regression model for clathrate formation dataset

Nithya S

2023-12-08

## Introduction

A clathrate is a chemical substance consisting of a lattice that traps or contains molecules. Clathrate hydrates are not officially chemical compounds, as the enclathrated guest molecules are never bonded to the lattice. The formation and decomposition of clathrate hydrates are first order phase transitions, not chemical reactions. The dataset considered here is about the Clathrate formation.

## Objective

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 multi-collinearity 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?

**URL for the dataset:**

*https://docs.google.com/spreadsheets/d/1WYlGKzpvfHr6PIgxgyPlo_1E6qYRJk5F/edit?usp=sharing&ouid=110138716074493614410&rtpof=true&sd=true*

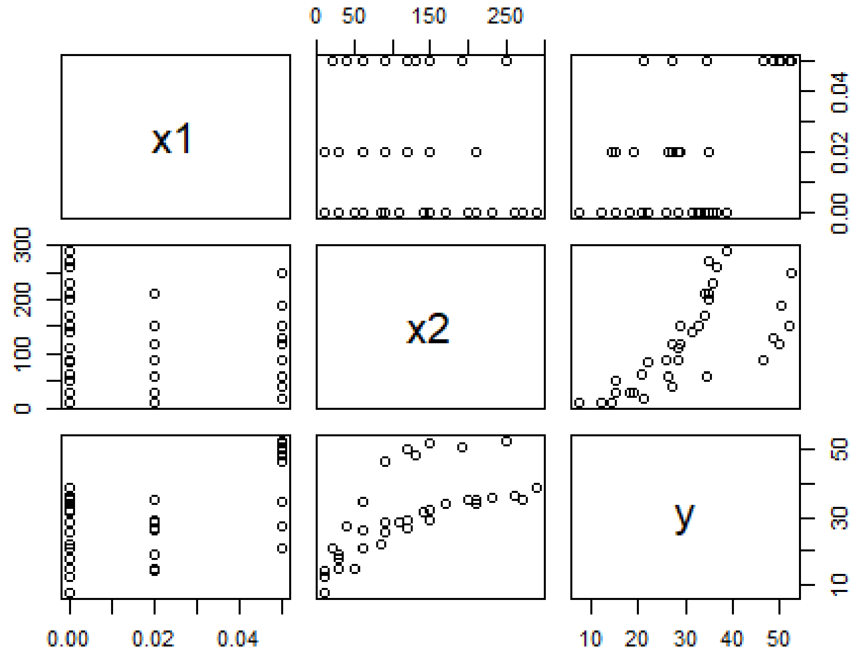
First we import the dataset as follows:

```
library(readxl)
Cathedral_data_formation <- read_excel("G:/My Drive/Linear
Regression/Datasets/Cathedral data-formation.xlsx")
View(Cathedral_data_formation)
attach(Cathedral_data_formation)
```

### 1. Fit a suitable linear regression model.

The suitable linear regression model for the dataset would be “Multiple regression model”. In order to fit the model, we first plot the variables to understand the relationship between them. Since there are two independent and one dependent variables, we plot the data using pairs as follows.

```
pairs(Cathedral_data_formation[1:3])
```



The correlation matrix for the same would be

```
cor(Cathedral_data_formation)

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

From the above matrix, we see that the correlation between $X_1$ and $X_2$ is -0.1275387, which says that the correlation between the two is very low. The correlation between Y and $X_1$ is 0.5192537, which says that the relationship is positive moderate. Similarly, the correlation between Y and $X_2$ is 0.6838246 which is again a positive moderate correlation.

Now we fit the model as follows:

```
model_cathedral_data_formation=lm(y~.,data=Cathedral_data_formation)
model_cathedral_data_formation

##
## Call:
## lm(formula = y ~ ., data = Cathedral_data_formation)
##
## Coefficients:
## (Intercept)           x1            x2
##     11.0870     350.1192        0.1089
```

```
summary(model_cathedral_data_formation)
```

```
##
## Call:
## lm(formula = y ~ ., data = Cathedral_data_formation)
##
## 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
```

The model built is

$$Y = B_0 + B_1X_1 + B_2X_2 + E$$

where, Y is the independent variable, $X_1$ and $X_2$ are the dependent variables, $B_0$, $B_1$ and $B_2$ are the intercept and slope regression coefficients and, E is the error associated in the model

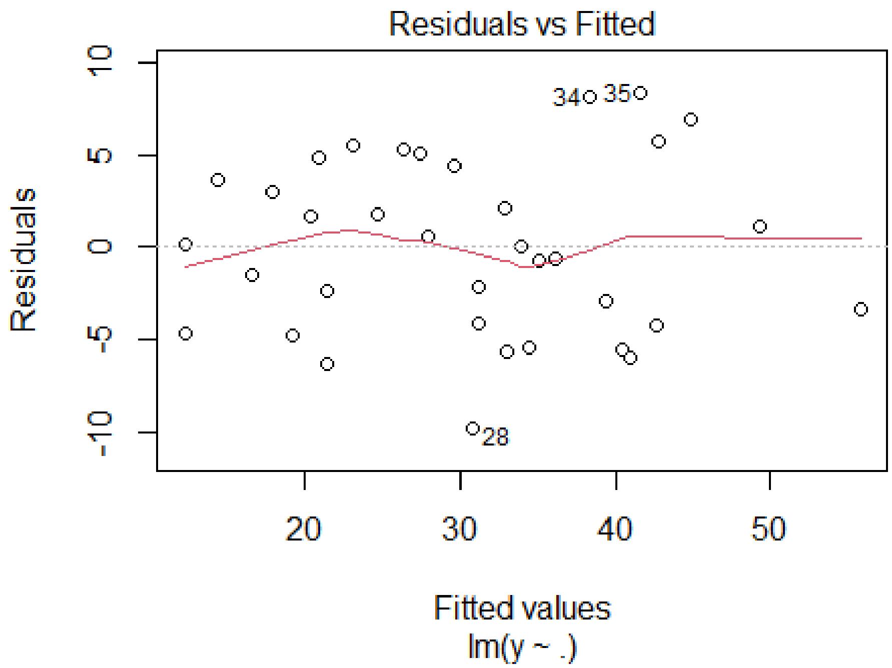
After the estimation of the model, we have

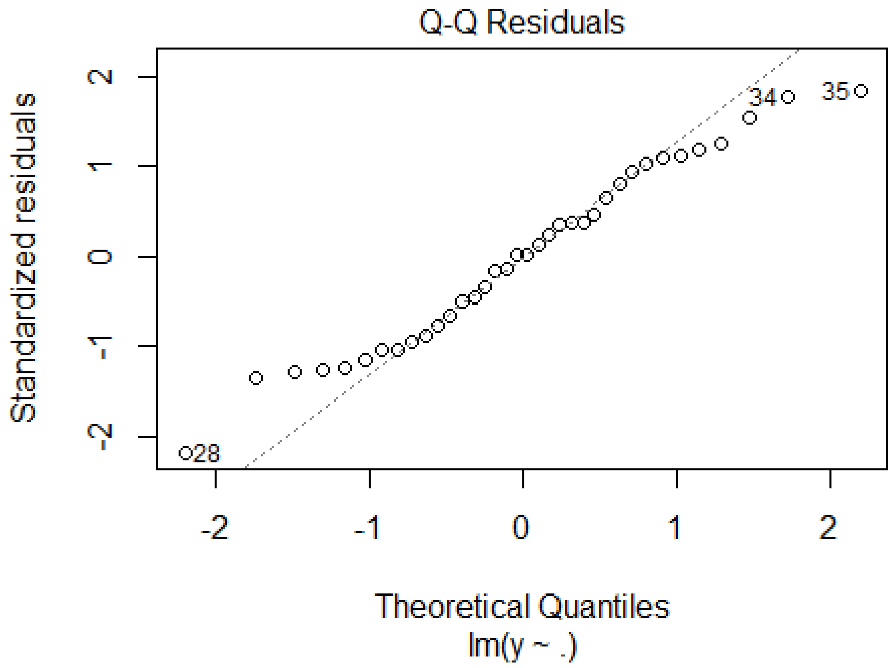
$$Y^{hat} = 11.0870 + 350.1192X_1 + 0.1089X_2$$

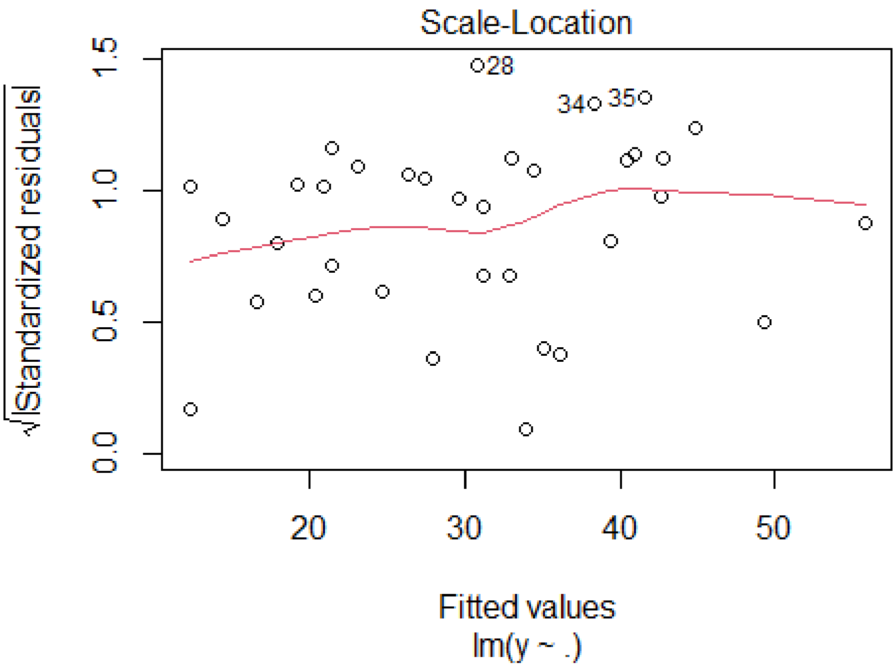
We infer from the above summary of the model as the average increase in Y when both $B_1$ and $B_2$ are zero is 11.0870. When $B_0$ is zero and when $X_2$ is fixed the unit increase in $X_1$ will result an increase in Y by 350.1192. Similarly, when $X_1$ is fixed and $B_0$ is zero, unit increase in $X_2$ will increase Y by 0.1089. Also we infer from the p-values in the table as, all the regression coefficients are significant. Also 83.19% of the explanation is given by the explanatory variables for the response variable.

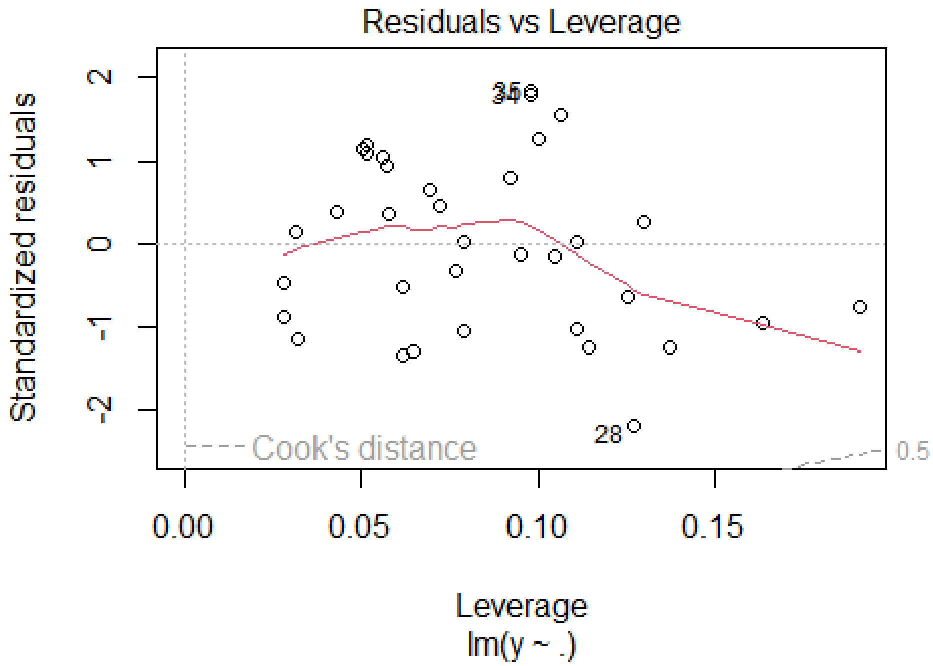
### 2. Construct a normal probability plot of the residuals. Does there seem to be any problem with the normality and constant variance assumption? If yes, what remedial measure will u perform?

```
plot(model_cathedral_data_formation)
```


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


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





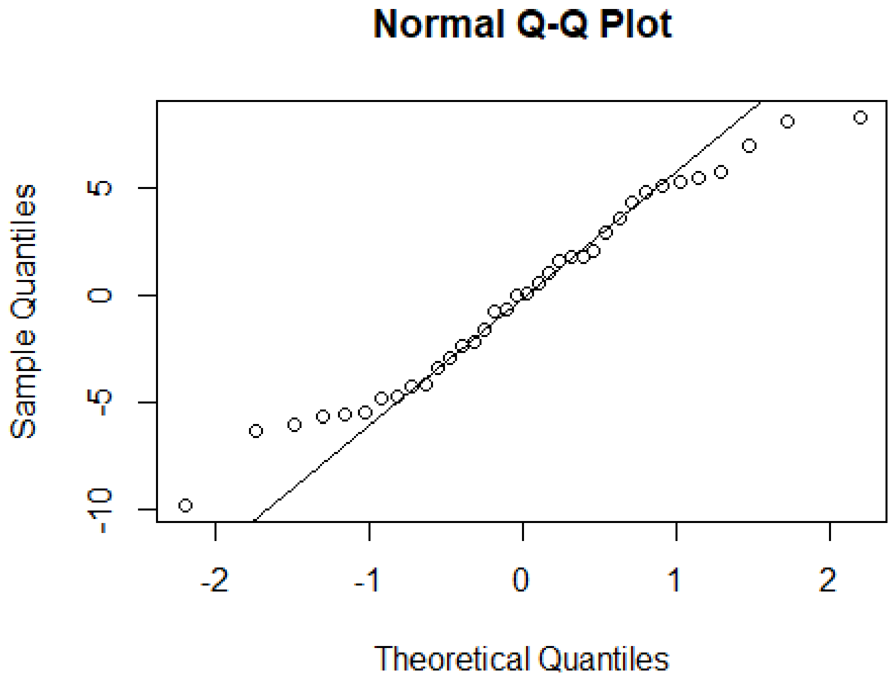
Assumption of normality:

1. QQ-plot

2. PP-plot

3. Kolmogorov/Shapiro test

```
residual=resid(model_cathedral_data_formation)
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

qqnorm(residual) # QQ plot-of residual

qqline(residual) # plots the points
```



If not all, majority of the points fall on the line thus the quartile of normal and residual are almost same, hence it indicates that the residuals follow a normal distribution. However, the assumption of normality has to be verified by using a statistical test. (But to know if the deviation of the points are lying away from the line, we use the test to further confirm the normality.)

Hypothesis to test for normality:

**$H_0$: Errors follow normal distribution.**

**v/s**

**$H_1$: Errors do not follow normal distribution.**

Normality using Shapiro

```
shapiro.test(residual)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residual
## W = 0.97094, p-value = 0.4519
```

At 0.05 level of significance the p value is 0.4519 which is greater than 0.05 thus we fail to reject null and say that the residuals follow normal distribution. Hence the assumption of errors.

Hypothesis testing for constant variance:

To test if the errors have constant variance.

**$H_0$: Errors have constant variance**

**v/s**

**$H_1$: Errors have no constant variance**

```
library(lmtest)
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

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

```
bptest(model_cathedral_data_formation)
```

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

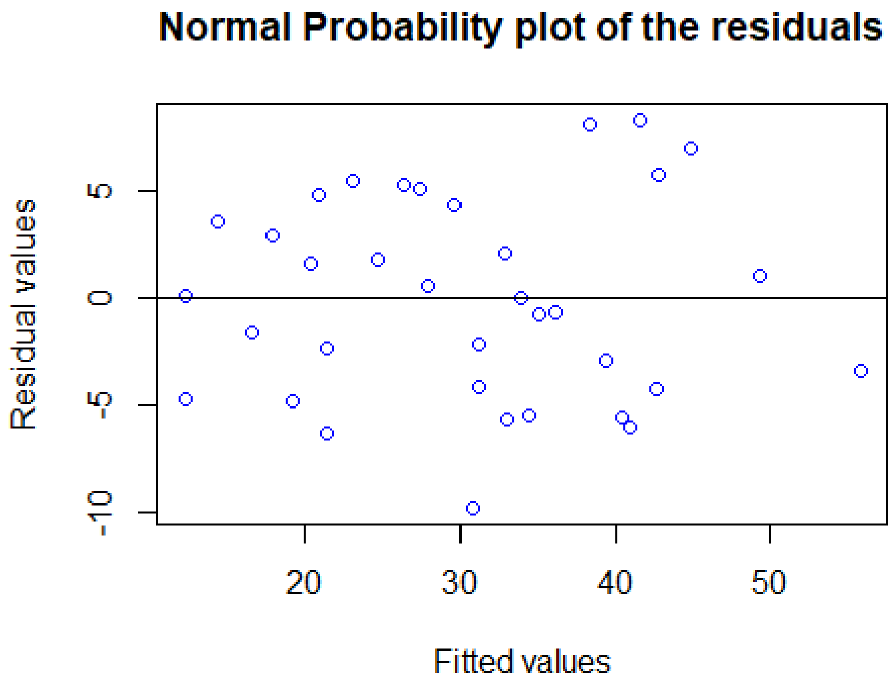
Since p-value is 0.01747 which is lesser than 0.05, thus we reject Ho and say that the errors have no constant variance. Hence the assumption of constant variance is not validated.

There is no problem with normality as all errors follow a normal distribution. But the assumption of constant variance is not validated. We can fix this problem by performing a transformation. In our model, we can take the logarithm transformation. Then we can refit our model and proceed for further analysis. By doing this transformation, the issue of constant variance can be fixed.

### 3. Construct and interpret a plot of the residuals.

To plot the residuals, we first get the fitted values for the model. Then we plot the fitted values and the residual values as follows:

```
fit=fitted.values(model_cathedral_data_formation)
fit

##        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

residual=resid(model_cathedral_data_formation)
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

plot(fit,residual,col="blue",main="Normal Probability plot of the
residuals",xlab="Fitted values",ylab="Residual values")

abline(0,0)
```



From the plot, we say that, the residuals are in a horizontal band fashion and they fluctuate more or less in a random manner inside, then this indicates that there are no visible model defects. So no conclusion can be drawn based on the linearity of the variables and constant variance. However we have shown above that the errors have no constant variance.

### 4. Are the residuals correlated?

To check if the residuals are uncorrelated, we can check for autocorrelation in the residuals by two ways,

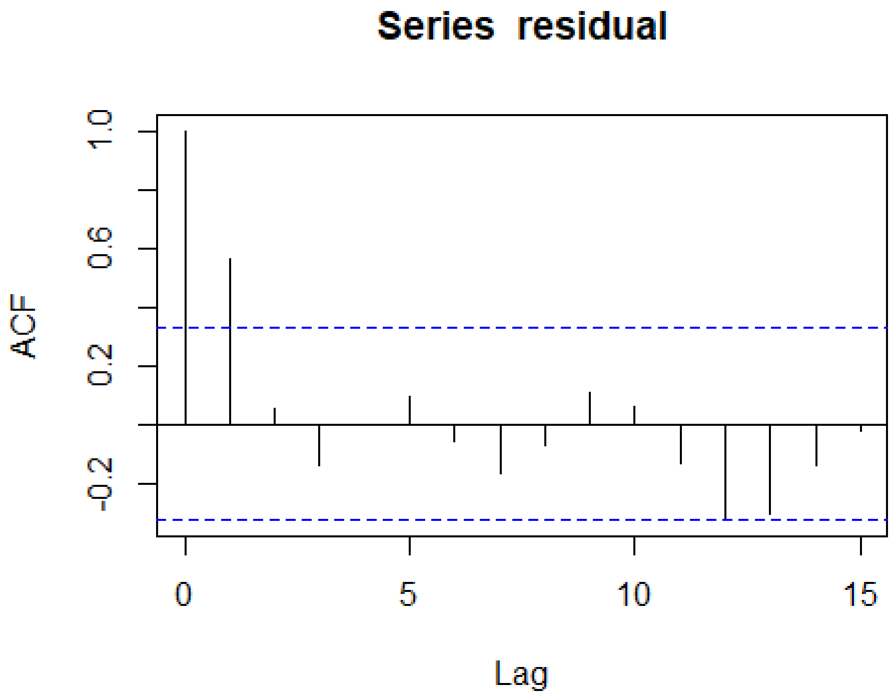
1. By plotting the ACF [Autocorrelation Function] curve

2. By performing the Durbin Watson test.

First, we plot the ACF curve:

```
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
```

```
ACF(residual)
```



From the graph we say that all autocorrelation values fall within the threshold limit (except one point). Since almost all values fall in the range from lag 1, it indicates that there is no significant autocorrelation among the residual series.

For examining the Durbin Watson test, the hypothesis to be tested:

**$H_0$: No autocorrelation**

**v/s**

**$H_1$: Autocorrelation**

```
dwtest(model_cathedral_data_formation)
```

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

At 5% significance level p-value 6.004e-06 is lesser than the significance level 0.05. So we reject the null hypothesis that there is no autocorrelation between the residual series.

Hence to summarize, we conclude that the residuals have no autocorrelation, i.e the residuals are not correlated.

### 5. Is multi-collinearity a potential problem in your model? If it is a problem, what is your remedy?

To check for multi-collinearity, we check if there is correlation greater than 0.7 for the independent variables, here it is X1 and X2. To check this, we call see the correlation of the model

```
cor(Cathedral_data_formation)

##             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 see that, there is no correlation between $X_1$ and $X_2$ that is greater than 0.7 So we conclude that there is no multi-collinearity in the considered model. If there exists the problem of multi-collinearity, we can fix it as follows:

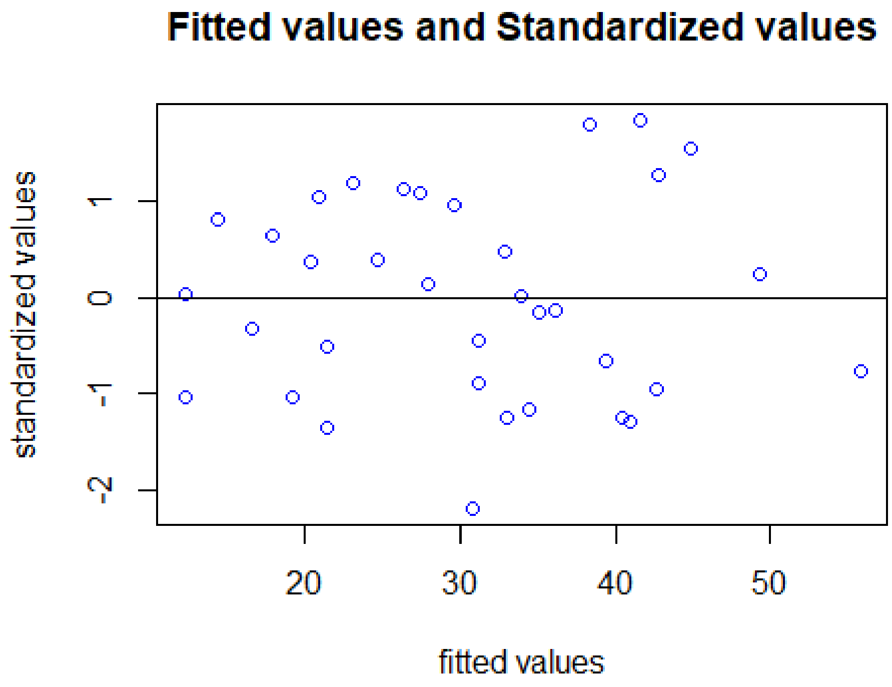
We first check for the correlation among the variables greater than 0.7, if there are any we compute the “Variance Inflation Factor”. If the VIF is greater than 5, we conclude that the corresponding variable is correlated with the any other variable. Hence we need to remove the variable and proceed for re-fitting.

We also need to understand that the VIF throws light on how the variance of the regression coefficients are highly affecting the model due to presence of multicollinearity.

### 6. Are there any outliers in the data? If it exists, how will you treat it?

To check if there are any outliers and scale out the residual values we use standardized residuals.

```
plot(fit,rstandard(model_cathedral_data_formation),col="blue",main="Fitted
values and Standardized values",xlab="fitted values",ylab="standardized
values")
abline(0,0)
```

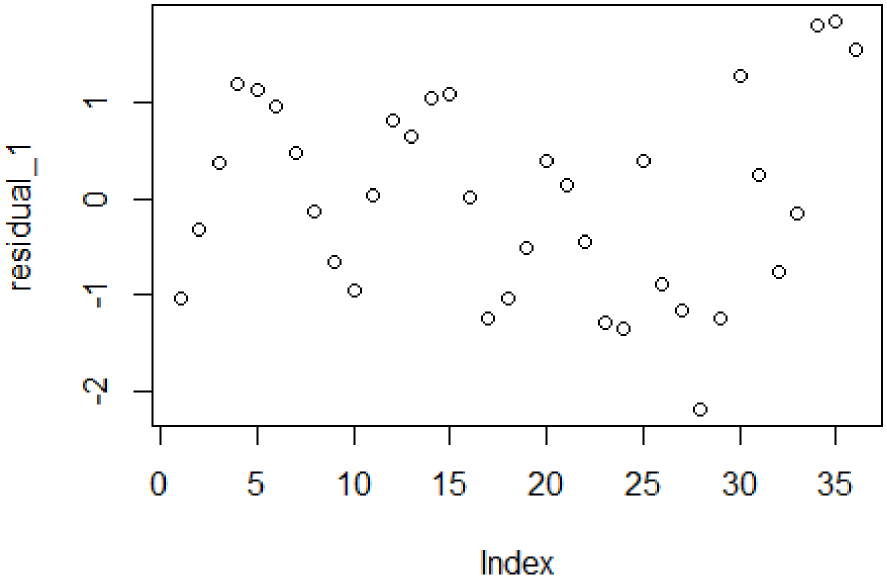


From the above plot we say that no values lieing outside the range of modulus 3. So there are no outliers in the dataset.

To be precise we can even check in the residual values for outliers as follows:

```
residual_1=rstandard(model_cathedral_data_formation)
residual_1

##           1            2            3            4            5 
6 
## -1.037115175 -0.333758043  0.356275584  1.187266673  1.129176925 
0.946369325 
##           7            8            9           10           11 
12 
##  0.461534357 -0.141070312 -0.650555446 -0.955185630  0.027428737 
0.799919557 
##          13           14           15           16           17 
18 
##  0.641316052  1.035126037  1.089371414  0.008015465 -1.237884530 -
1.041331230 
##          19           20           21           22           23 
24 
## -0.508984388  0.379286883  0.128876651 -0.458446375 -1.289781318 -
1.351031103 
##          25           26           27           28           29 
30 
##  0.379286883 -0.882640059 -1.154169443 -2.186886007 -1.255401518
```

```
1.266420897
##           31           32           33           34           35
36
##  0.248696768 -0.773278110 -0.161131924  1.783666308  1.835215764
1.541125067

plot(residual_1)
```



Now we have numerically seen that there are no outliers in the dataset. If there exists outliers in the dataset, we remove the outliers from the dataset and proceed for the analysis by re-fitting the model.

## Conclusion

To summarize, we fitted the model for Clathrate formation dataset, and got the estimated model as,

$$Y^{hat} = 11.0870 + 350.1192X_1 + 0.1089X_2$$

We constructed a normal probability plot of the residuals. We saw that the errors followed normal distribution, however the assumption of constant variance was not satisfied. For the model considered here, this issue can be resolved by taking logarithm and re-fitting the model.

We constructed and interpreted the plot of the residuals. We inferred from the graph as, the residuals are in a horizontal band fashion and they fluctuate more or less in a random manner inside, then this indicates that there are no visible model defects. So no conclusion can be drawn based on the linearity of the variables and constant variance. However we have shown above that the errors have no constant variance.

We checked for the correlation in the residuals and found that, the residuals have autocorrelation. We checked this by using ACF plot and also Durbin Watson test.

There was no issue of multicollinearity in our model.

There were no outliers in our considered dataset.