Project 2

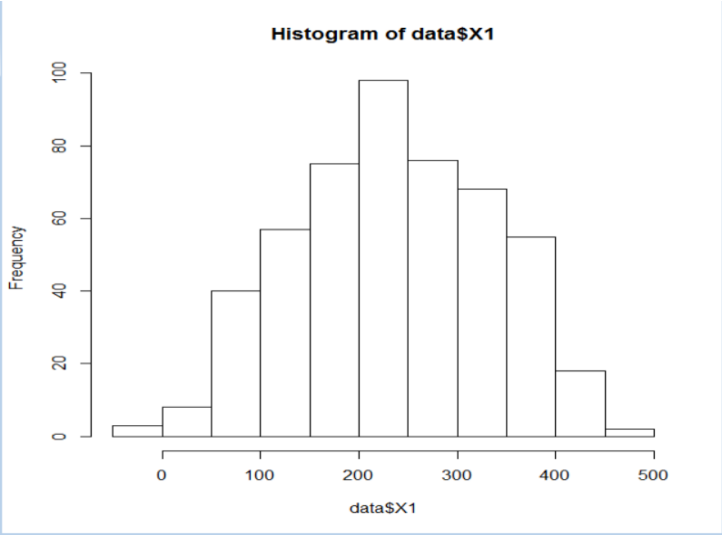
rpatel17

Ronak Dipankumar Patel

Ans 1.) Basic Statistics Analysis:

1.1)

Variable X1



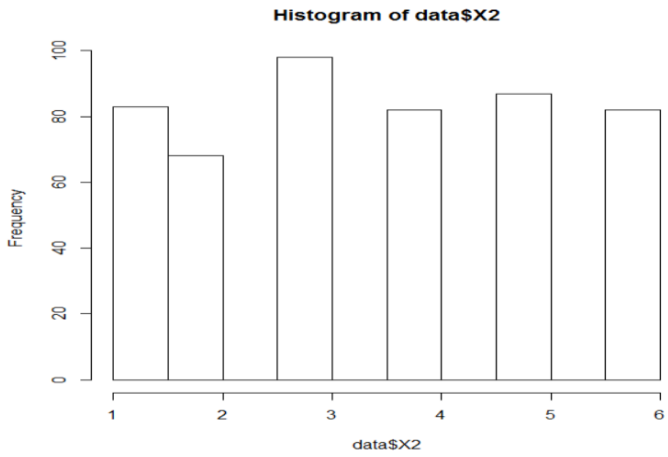
Mean :

[1] 235.6931

Variance :

[1] 9915.638

Variable X2



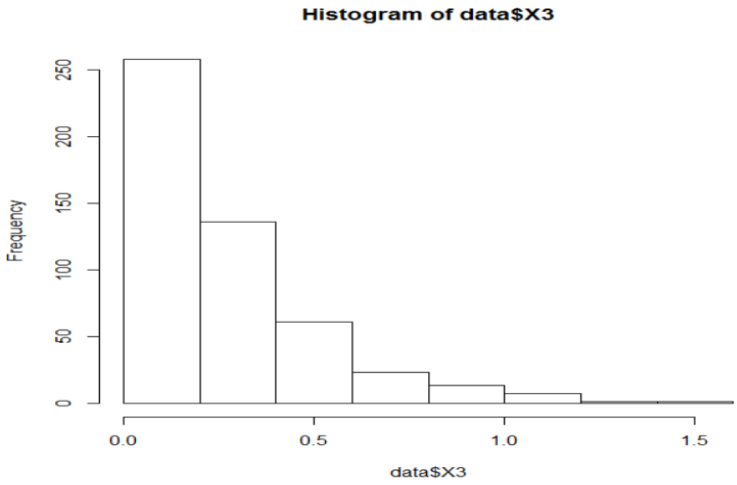
Mean :

[1] 3.536

Variance :

[1] 2.854413

Variable X3



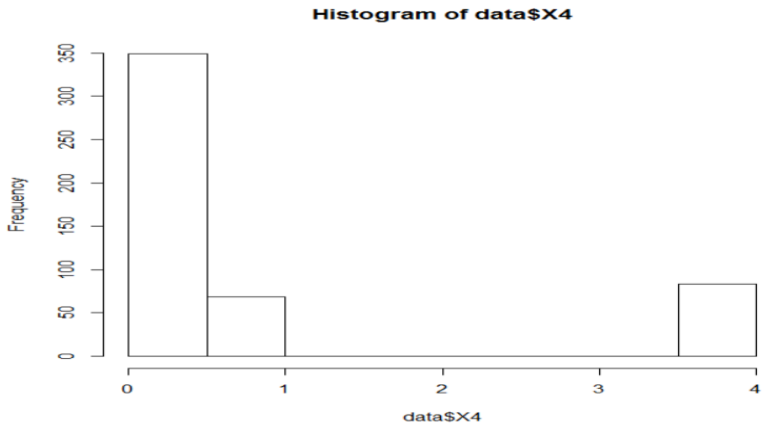
Mean :

[1] 0.2637452

Variance :

[1] 0.05882621

Variable X4



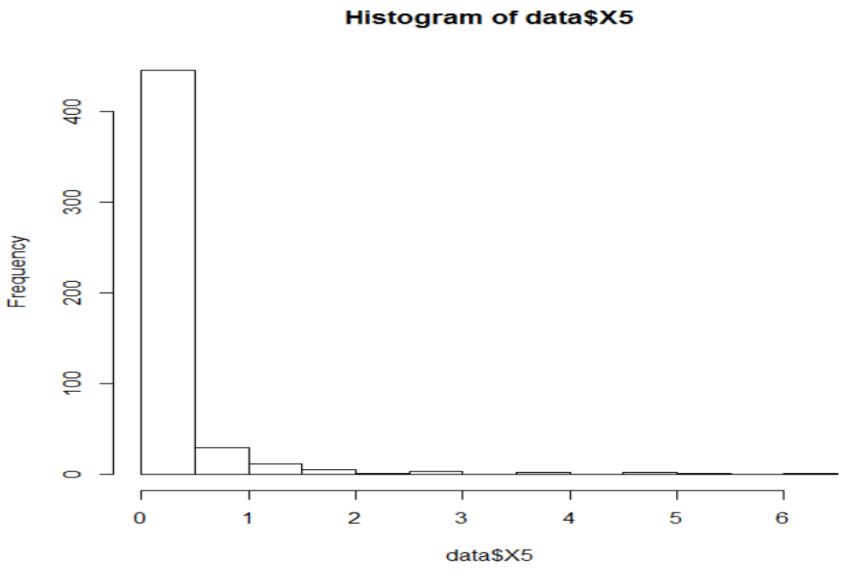
Mean :

[1] 0.9741733

Variance :

[1] 1.902236

Variable X5



Mean :

[1] 0.2205745

Variance :

[1] 0.3852349

1.2)

Interpreting Correlation amongst different variables:

	X1	X2	X3	X4	X5	Y
X1	1.00000000	0.092094622	0.011807992	-0.071318727	0.274340595	0.98651390
X2	0.09209462	1.000000000	0.043463697	-0.788987335	0.006093379	0.18096608
X3	0.01180799	0.043463697	1.000000000	-0.036716606	0.008255306	0.01948859
X4	-0.07131873	-0.788987335	-0.036716606	1.000000000	-0.001229658	-0.13453729
X5	0.27434059	0.006093379	0.008255306	-0.001229658	1.000000000	0.28911626
Y	0.98651390	0.180966080	0.019488591	-0.134537290	0.289116258	1.00000000

1.3)

Strength of Linear relationship:

	X1	X2	X3	X4	X5	Y
X1	Perfect	None to extremely weak	None to extremely weak	None to extremely weak	Weak	Very Strong Positive
X2	None to extremely weak	Perfect	None to extremely weak	Strong Negative	None to extremely weak	None to extremely weak
X3	None to extremely weak	None to extremely weak	Perfect	None to extremely weak	None to extremely weak	None to extremely weak
X4	None to extremely weak	Strong Negative	None to extremely weak	Perfect	None to extremely weak	None to extremely weak
X5	Weak	None to extremely weak	None to extremely weak	None to extremely weak	Perfect	Weak
Y	Very Strong Positive	None to extremely weak	None to extremely weak	None to extremely weak	Weak	Perfect

Clearly from the above table we could see that X1 and Y are very strongly correlated and also X2 and X4 have strong negative correlation. Thus, whenever we take multiple variables into our regression analysis we should avoid using both X2 and X4 as it would cause redundancy and make to model to underfit or overfit the data. If we were to take only one variable for our analysis, it would be X1.

Ans 2.)

2.1)

a_0 = 32.88083

a_1 = 1.45834

```
Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-75.838 -17.644   0.682  16.134  70.684

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 32.88083    2.77379   11.85   <2e-16 ***
x            1.45834    0.01084  134.50   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 24.12 on 498 degrees of freedom
Multiple R-squared:  0.9732,    Adjusted R-squared:  0.9732
F-statistic: 1.809e+04 on 1 and 498 DF,  p-value: < 2.2e-16
```

s^2 = 581.6727

```
[1] 581.6727
```

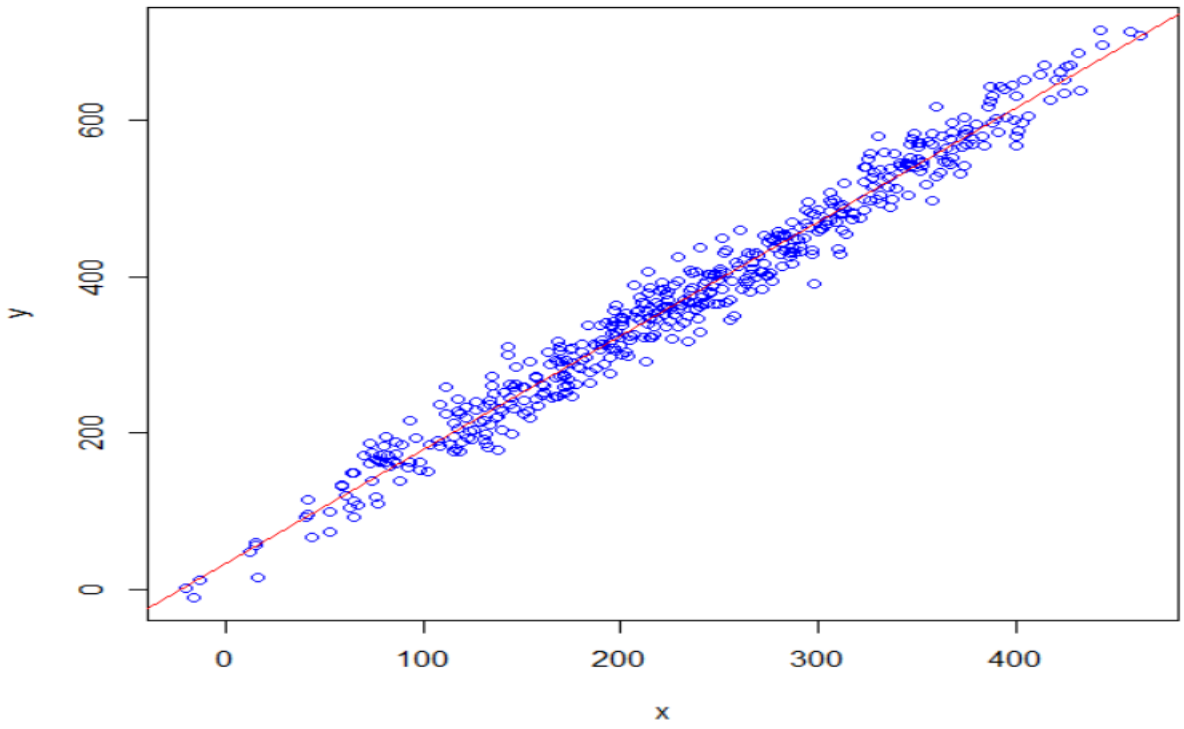
2.2)

The p-value for each term tests the null hypothesis that the coefficient is equal to zero. A low p-value for a_0 and a_1 (< 0.05) indicates that you can reject the null hypothesis. Thus, the p-values suggests that the coefficients are meaningful.

R^2 of 0.9732 implies that the model very well fits the data. However, R^2 cannot determine whether the coefficient estimates and predictions are biased. Also, in this instance, ~ 97% of the Y value is due to the X1 value.

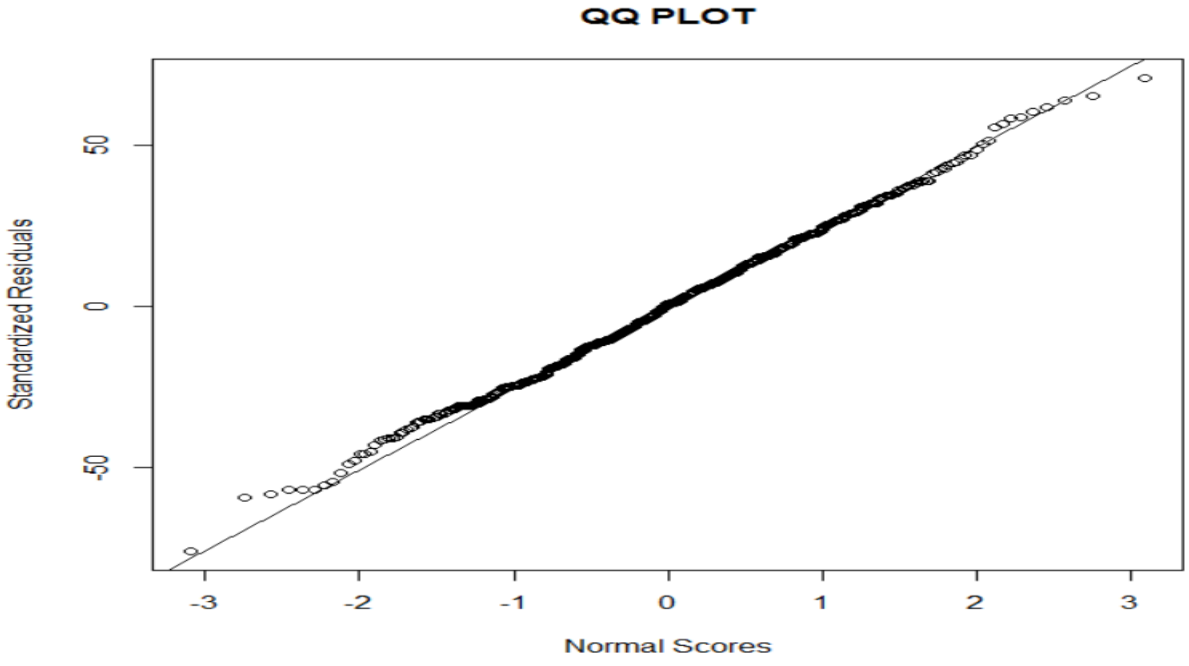
The F-test in our case seems to reject the null hypothesis since the p-value for the F-statistic is very small (null hypothesis: that the intercept only model and our model are equal).

2.3)

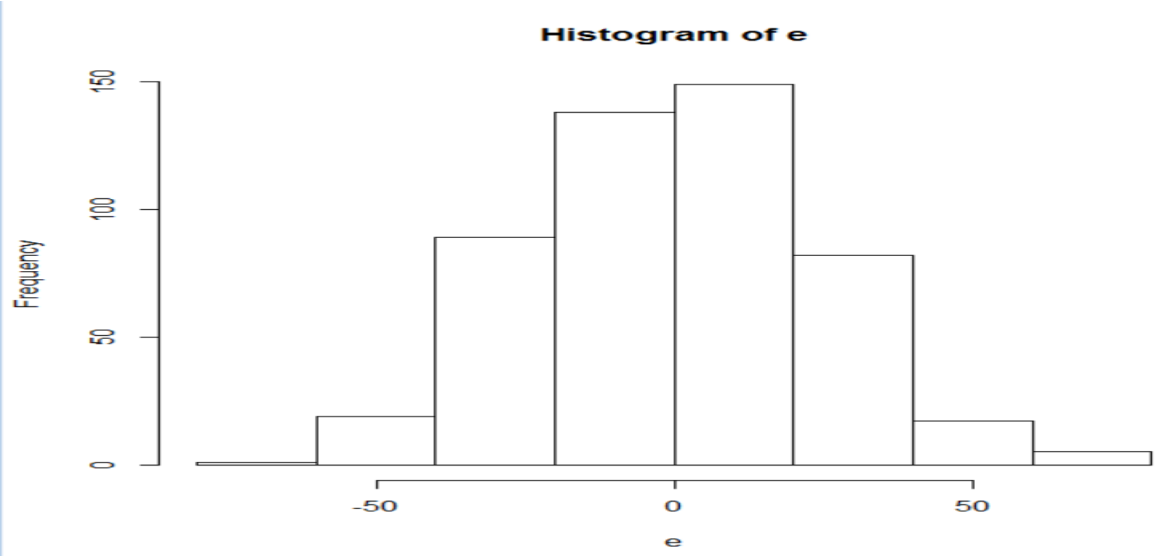

y
x
0
200
400
600
0
100
200
300
400

2.4)

a)

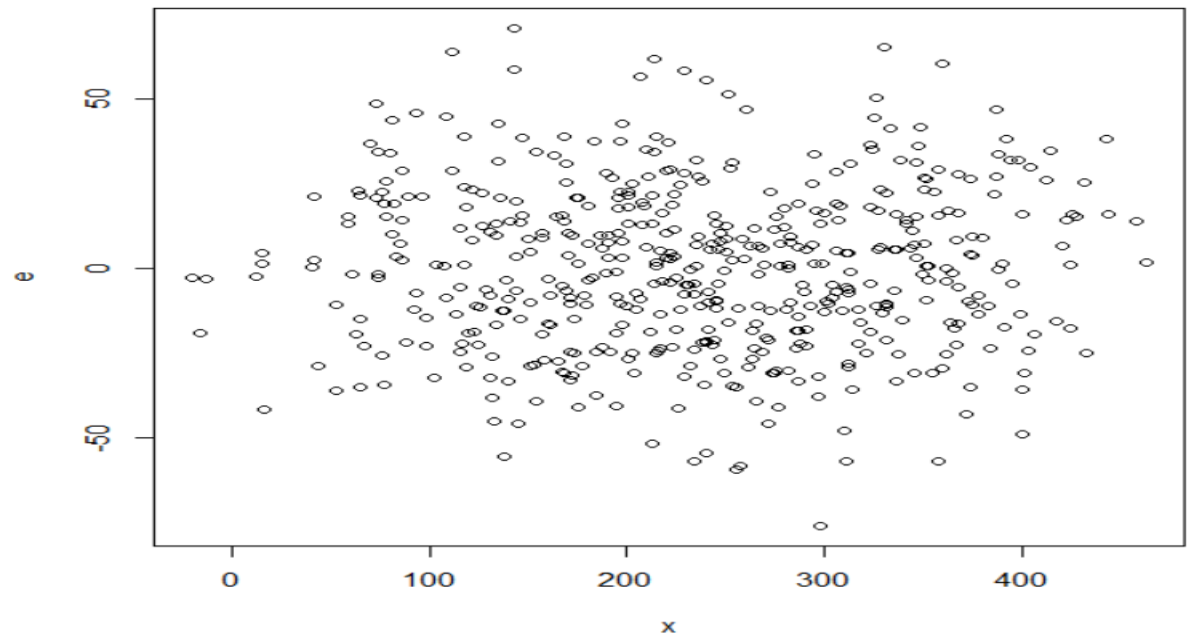

QQ PLOT
Standardized Residuals
Normal Scores
-50
0
50
-3
-2
-1
0
1
2
3

Clearly, the QQ plot shows us that except for the tails the residual points follow the straight line and imply that the observed residuals are normally distributed. The Symmetric distribution is presented in the histogram below:



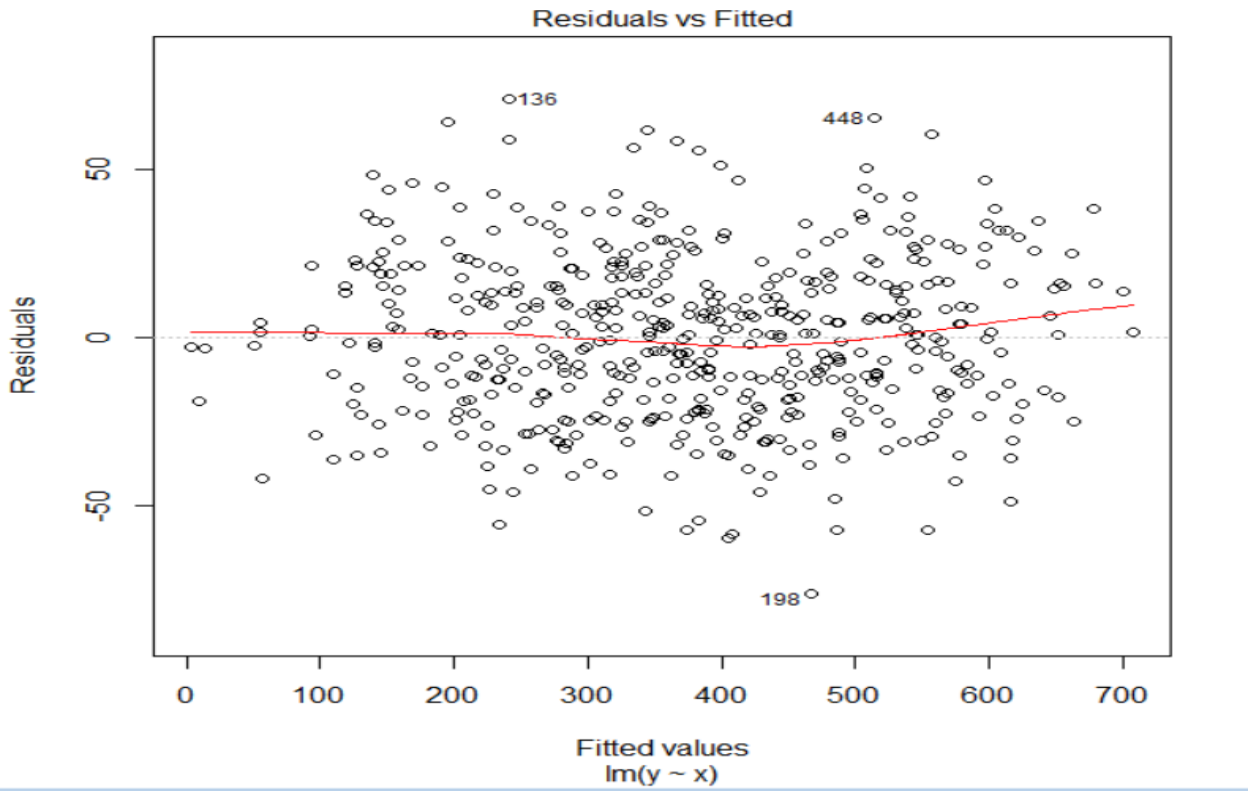
b.)

As you can clearly see below there are no correlation trends. The variability is constant throughout the plot. There isn't any curvature or other indications implying that there is a problem with our model.



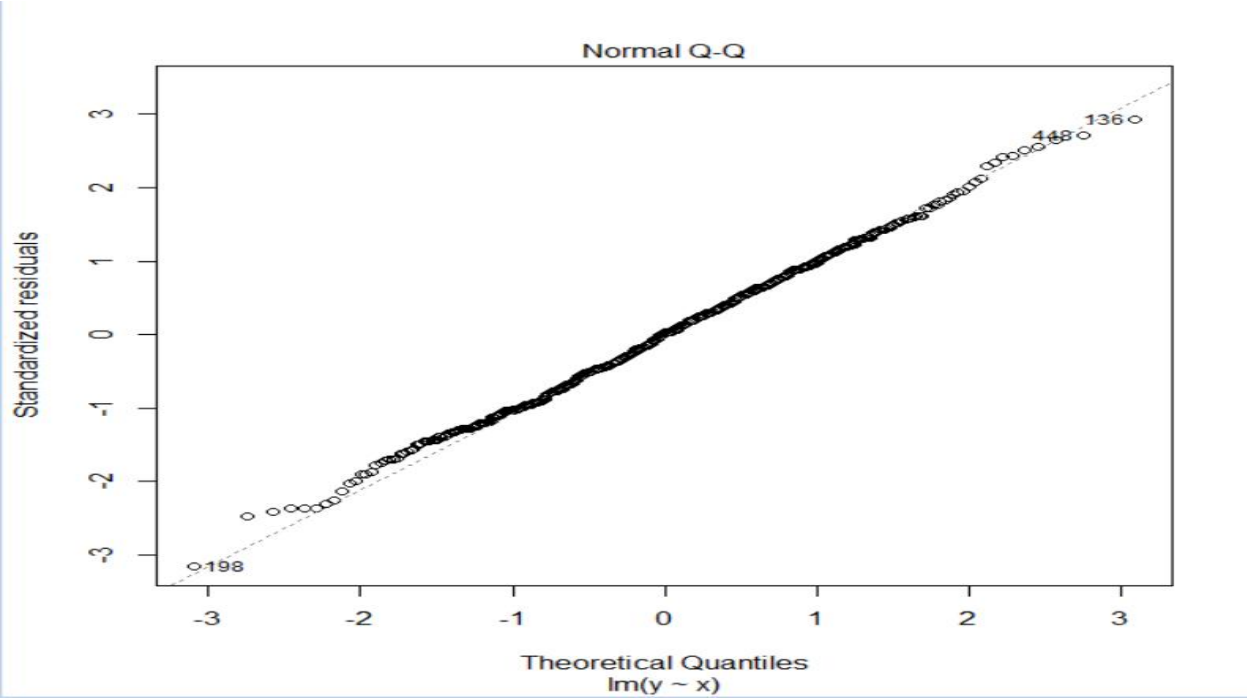
2.5) and 2.6)

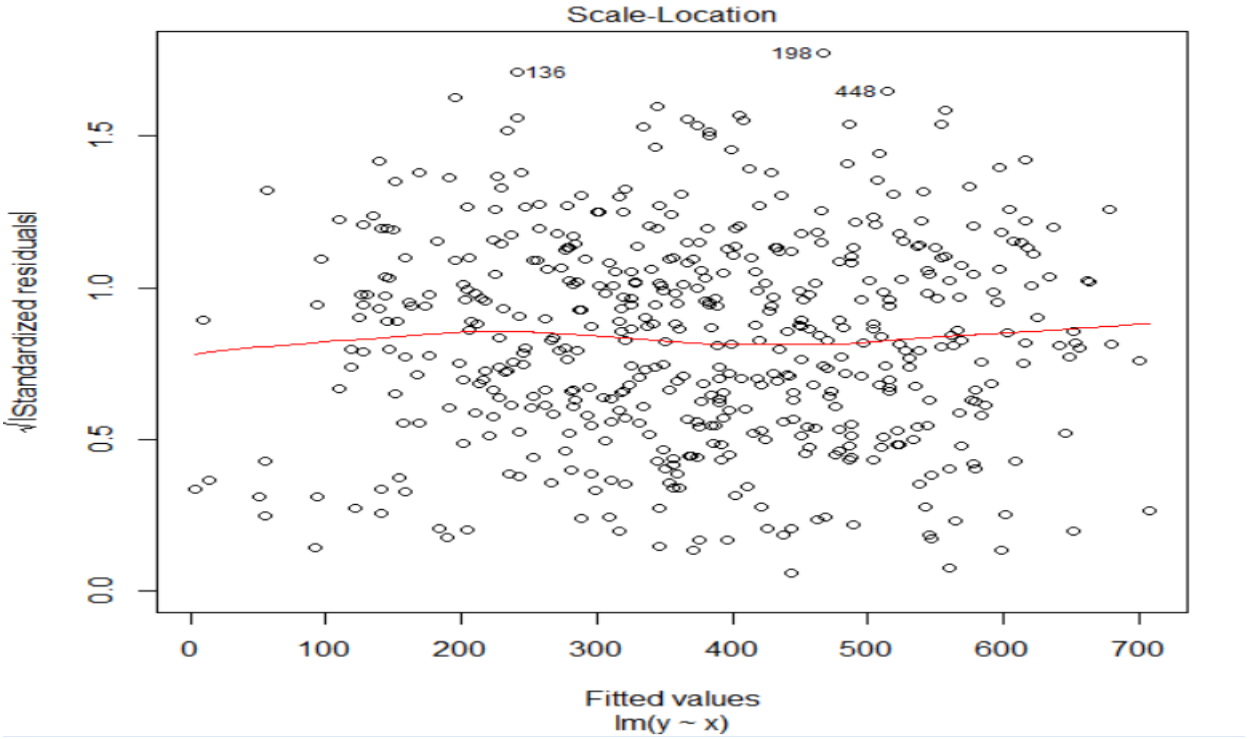
The below are the first two diagnostic plots available in R for the above fit:



The above plot gives us an idea whether there is curvature in the data. If the red line is curved strongly, a quadratic or other model is better. However in our case, the curvature is not so strong.

The below plot tells us whether the residuals are normally distributed. In our case they are normally distributed since they follow the straight line.


Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(y ~ x)
198
448
136
-3
-2
-1
0
1
2
3


Scale-Location
√|Standardized residuals|
Fitted values
lm(y ~ x)
136
198
448
0.0
0.5
1.0
1.5
0
100
200
300
400
500
600
700

The plot above is used to check if the variance is constant. It is used to check to see if there are any overly influential points. The variance appears constant since the red line is not tilted up/down strongly.

Higher order polynomial regression fit and conclusion:

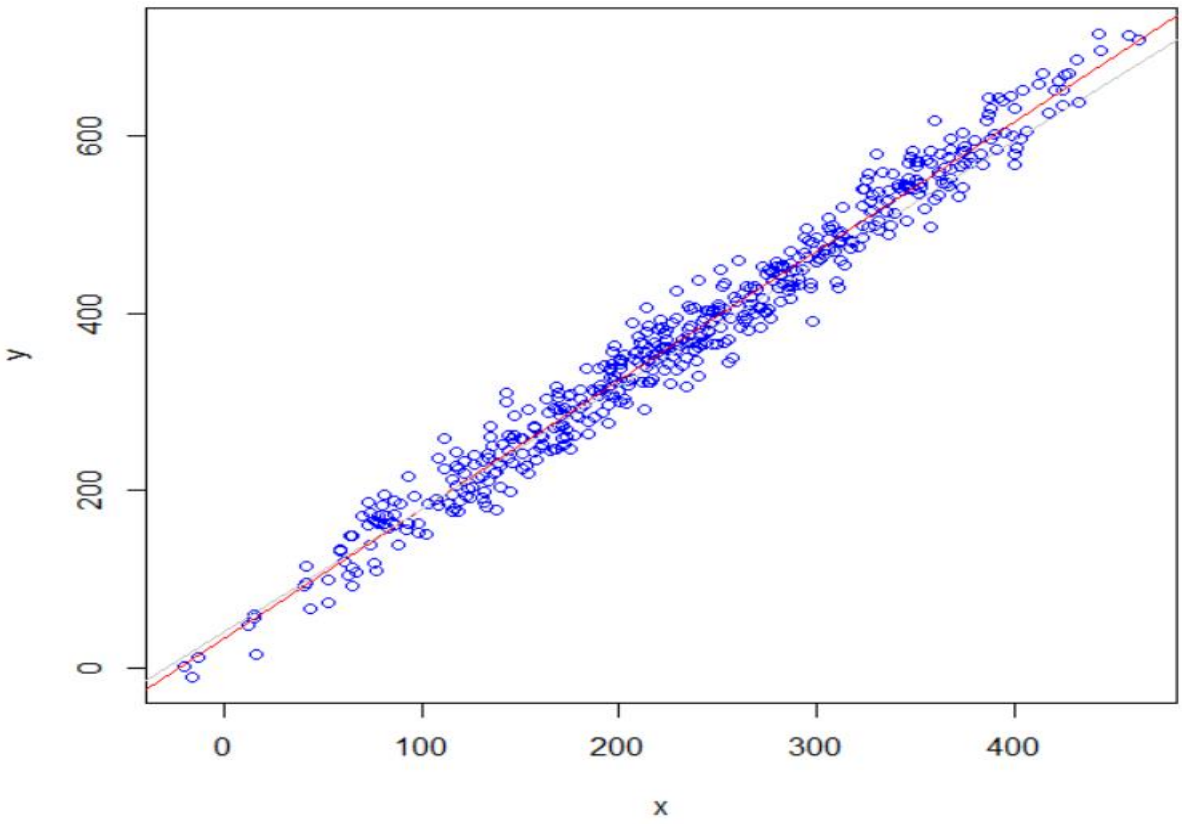
```
Call:
lm(formula = y ~ poly(x, 2, raw = T))

Residuals:
    Min      1Q  Median      3Q     Max
-75.003 -17.236  -0.152  16.350  70.967

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          3.939e+01  4.968e+00   7.930 1.46e-14 ***
poly(x, 2, raw = T)1 1.389e+00  4.494e-02  30.916  < 2e-16 ***
poly(x, 2, raw = T)2 1.486e-04  9.407e-05   1.579    0.115
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 24.08 on 497 degrees of freedom
Multiple R-squared:  0.9733,    Adjusted R-squared:  0.9732
F-statistic:  9074 on 2 and 497 DF,  p-value: < 2.2e-16
```

As seen earlier, there wasn't much curvature in the data. Otherwise, a quadratic or other model would have been better. Moreover, fitting in a quadratic model doesn't yield better results compared to the previous one. The Adjusted R-square value remain the same. Clearly, also the p-value for the x^2 coefficient is 0.115 which is greater than 0.05. Thus, the x^2 term is zero (in statistically terms) as there isn't enough evidence for the null hypothesis to be rejected.

Also, the variance obtained is `[1] 579.9334` .

Below is the plot for the higher order polynomial fit:

The red line is for the previous linear fit and the gray line is corresponding to the current quadratic fit :



Ans 3.)

3.1)

Using all the independent variable implies use either X2 or X4. I have used variable X1, X2, X3 and X5.

Summary:

$a_0 = 8.07$

$a_1 = 1.44$

$a_2 = 7.94$

$a_3 = 2.35$

$a_5 = 5.18$

```
Call:
lm(formula = y ~ x1 + x2 + x3 + x5)

Residuals:
    Min      1Q  Median      3Q     Max
-54.470 -13.314  -0.875  13.997  76.755

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 8.068231   2.980085   2.707 0.007016 **
x1          1.437000   0.009348 153.716  < 2e-16 ***
x2          7.940942   0.530327  14.974  < 2e-16 ***
x3          2.347899   3.678042   0.638 0.523537
x5          5.181530   1.493459   3.469 0.000567 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 19.91 on 495 degrees of freedom
Multiple R-squared:  0.9819,    Adjusted R-squared:  0.9817
F-statistic:  6697 on 4 and 495 DF,  p-value: < 2.2e-16
```

Sigma-square:

```
[1] 396.3152
```

3.2)

Clearly, after looking at the p-values we could remove attribute X3 from our analysis since the corresponding p-value is 0.523537 which implies that X3 is not statistically significant.

Clearly X1 and X2 have very low p-value and clearly they could not be left out. Also, X5 have a weak correlation with Y however, the p-value is <0.05 thus, it could not be considered zero.

Now, fitting in a regression line with variables X1, X2 and X5.

```
Call:
lm(formula = y ~ x1 + x2 + x5)

Residuals:
    Min      1Q  Median      3Q     Max
-53.903 -13.671  -0.674  13.746  77.000

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 8.626856   2.847001   3.030 0.002572 **
x1          1.437035   0.009343 153.815  < 2e-16 ***
x2          7.955389   0.529528  15.024  < 2e-16 ***
x5          5.187322   1.492539   3.476 0.000555 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 19.9 on 496 degrees of freedom
Multiple R-squared:  0.9818,    Adjusted R-squared:  0.9817
F-statistic:  8940 on 3 and 496 DF,  p-value: < 2.2e-16
```

```
[1] 395.8418
```

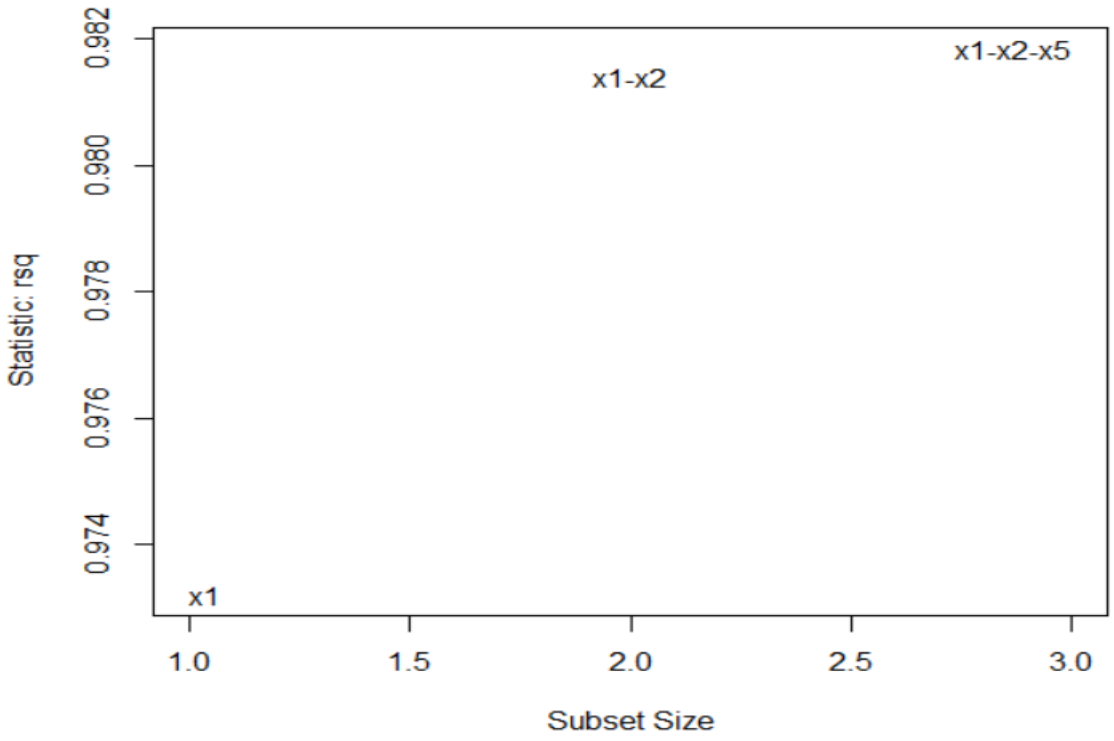
Now, fitting in a regression line with variables X1, X2.

```
Call:
lm(formula = y ~ x1 + x2)

Residuals:
    Min      1Q  Median      3Q     Max
-54.994 -13.480  -0.618  14.008  77.575

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 7.797114   2.868420   2.718  0.00679 **
x1          1.445963   0.009082 159.210  < 2e-16 ***
x2          7.918541   0.535290  14.793  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 20.12 on 497 degrees of freedom
Multiple R-squared:  0.9814,    Adjusted R-squared:  0.9813
F-statistic: 1.311e+04 on 2 and 497 DF,  p-value: < 2.2e-16
```

```
[1] 404.6659
```

So, if we carefully look at the adjusted R-square value for both the above models, we would realize that model with X1, X2 and X5 is better than model with only X1 and X2, which is better than model with only X1.

Here is the plot summing up the Adjusted R-square comparison of the three different models:



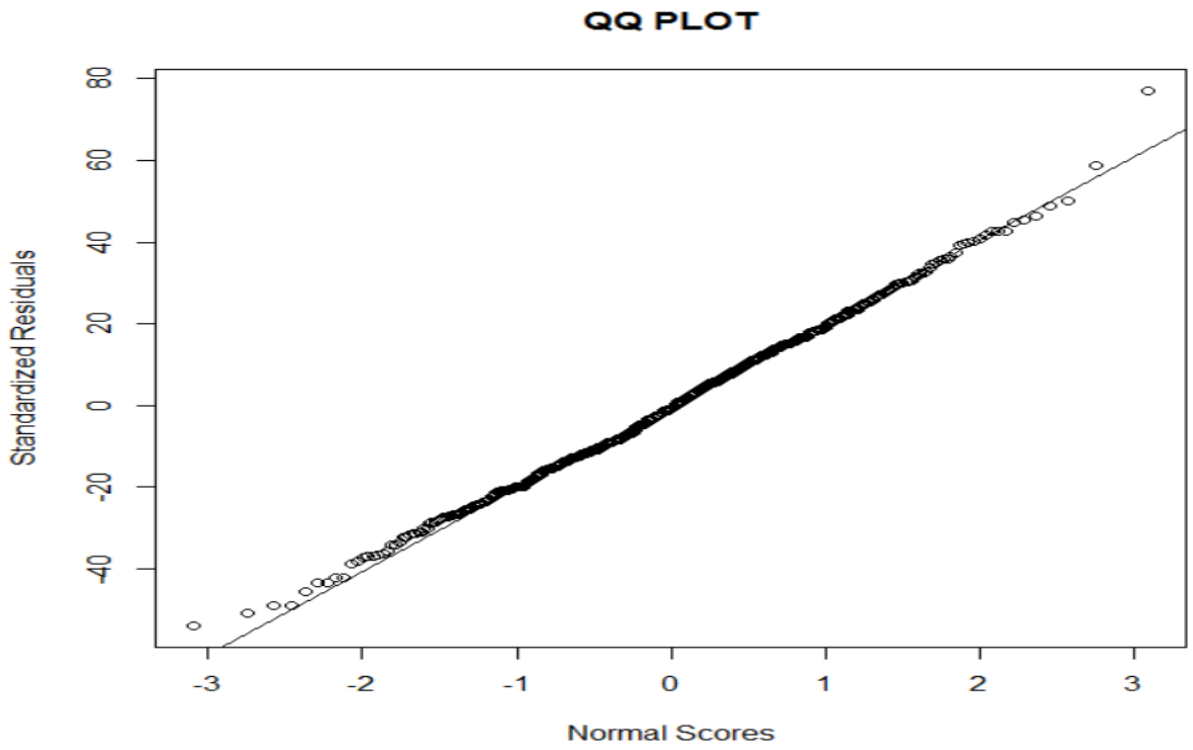
Also, if we check the summary for the model with variables X1, X2, X3 and X5: the adjusted R-square is 0.9817 and the summary for the model comprising all the variables have adjusted R-square value of 0.9818. However, the p-values and the correlation table tend to suggest that variables X3 and X4 need not be a part of the final model.

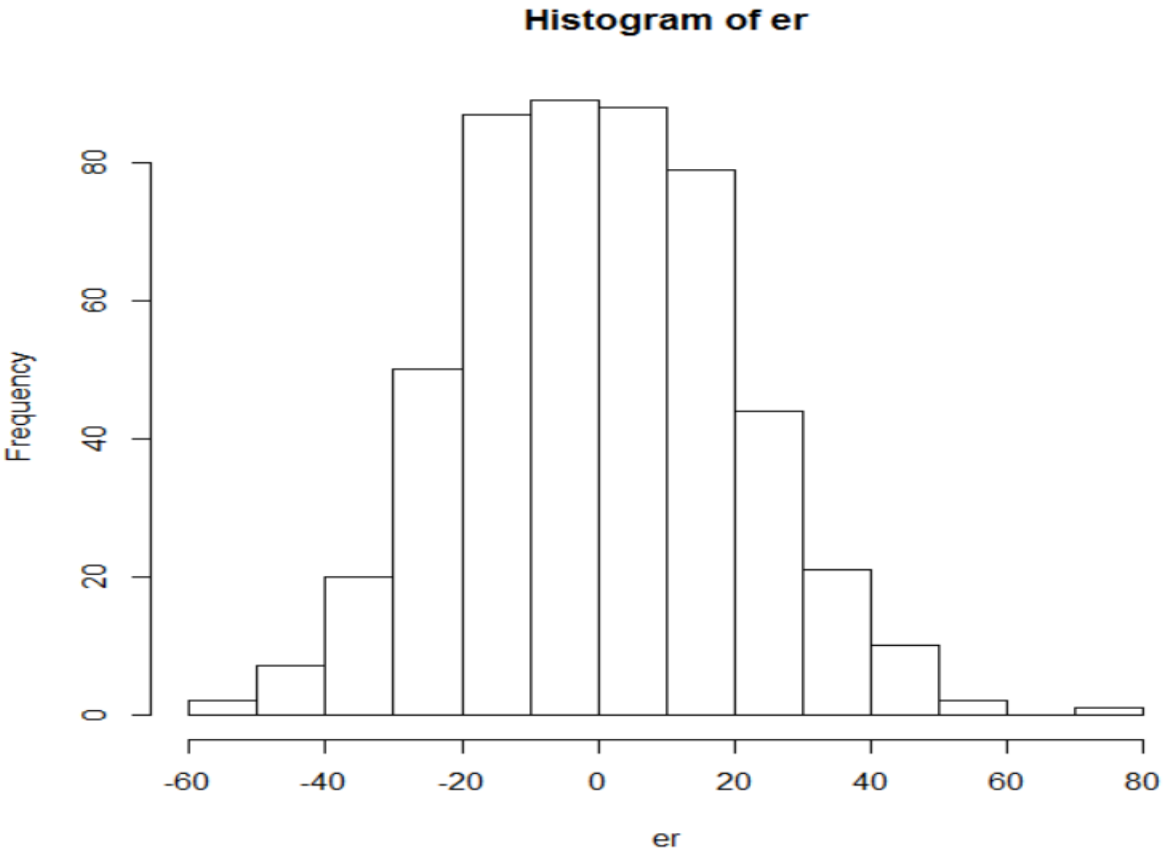
I have thus performed residual analysis for models with variables : X1, X2 and X5 and X1 and X2 respectively.

3.3) QQ plots and scatter plots:

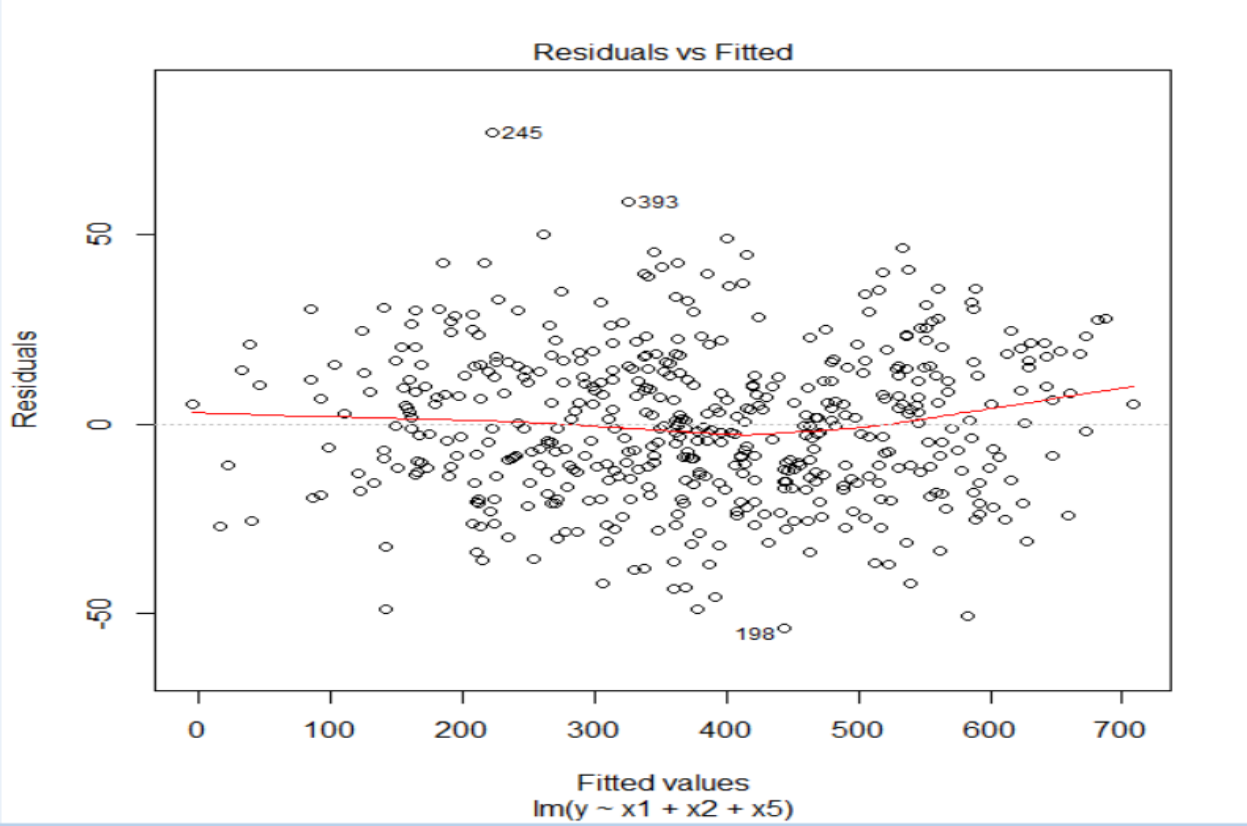
For model with variable X1, X2 and X5:

Clearly, the QQ plot shows us that except for the tails the residual points follow the straight line and imply that the observed residuals are normally distributed.


QQ PLOT
Standardized Residuals
Normal Scores
-40
-20
0
20
40
60
80
-3
-2
-1
0
1
2
3

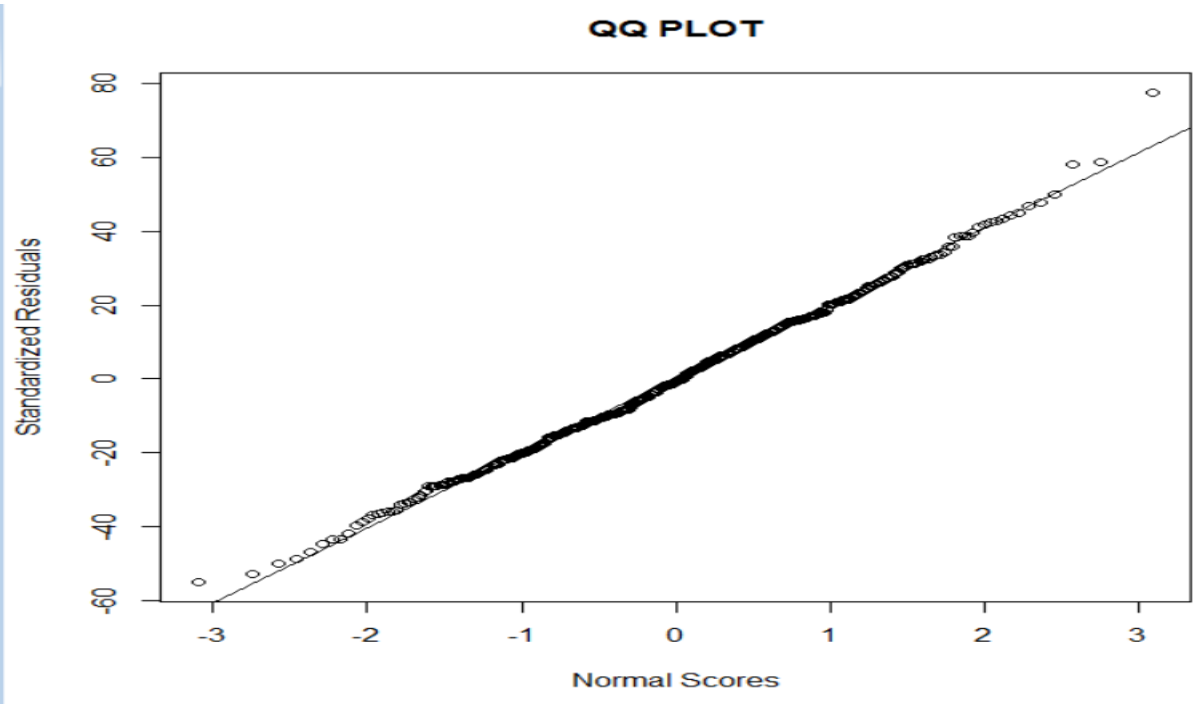

Histogram of er
Frequency
er
0
20
40
60
80
-60
-40
-20
0
20
40
60
80

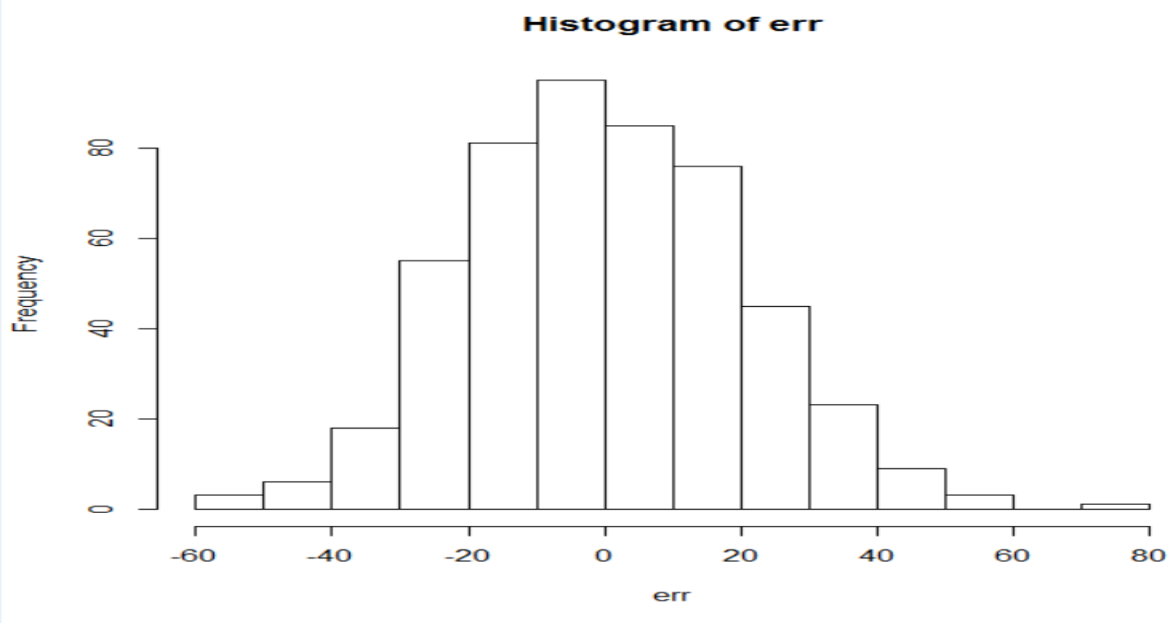
As you can clearly see below there are no correlation trends. The variability is constant throughout the plot. There isn't any curvature or other indications implying that there is a problem with our model.



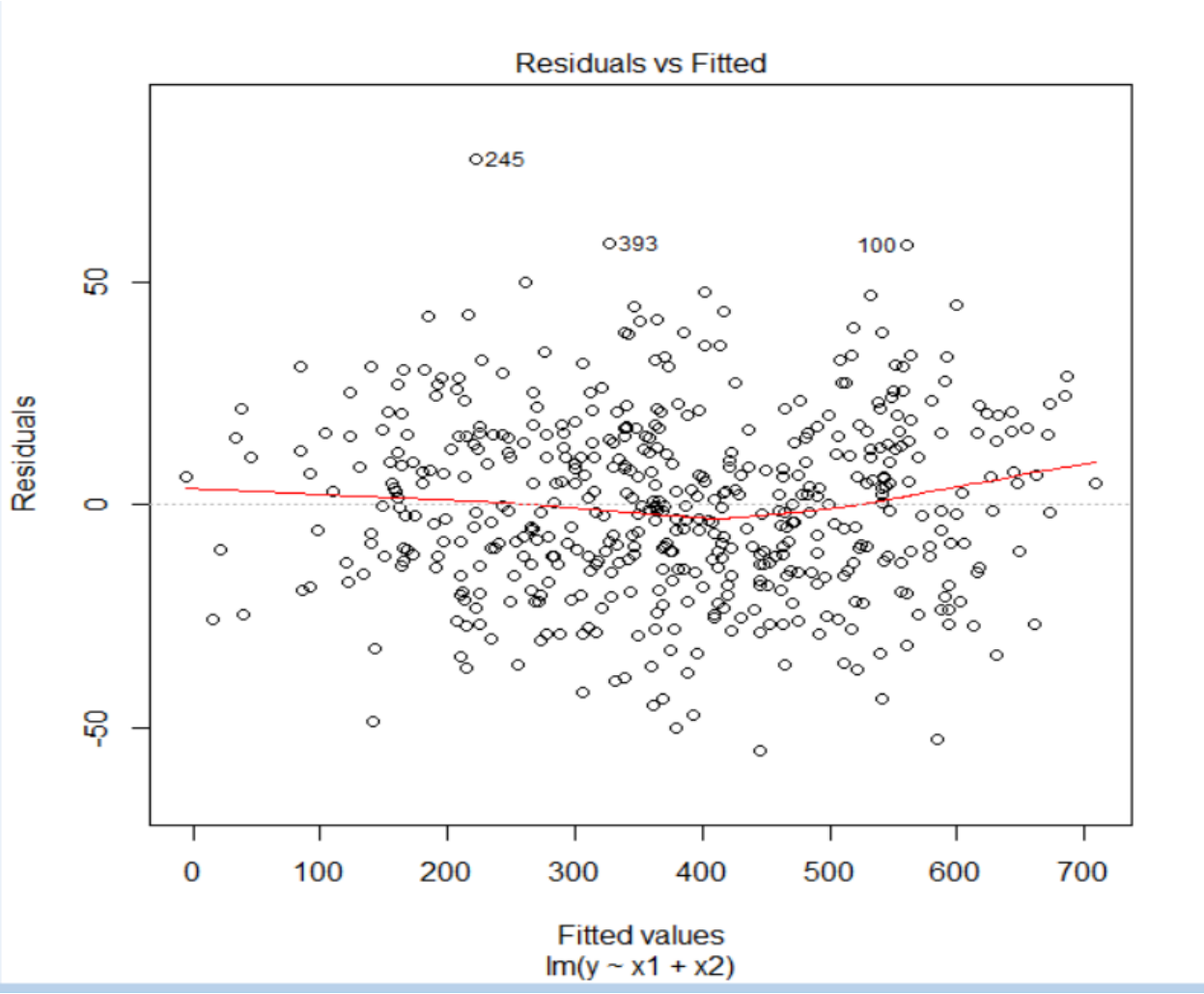
For model with variable X1, X2:

Clearly, the QQ plot shows us that except for the tails the residual points follow the straight line and imply that the observed residuals are normally distributed.





As you can clearly see below there are no correlation trends. The variability is constant throughout the plot. There isn't any curvature or other indications implying that there is a problem with our model.


Residuals vs Fitted
245
393
100
Residuals
50
0
-50
0
100
200
300
400
500
600
700
Fitted values
lm(y ~ x1 + x2)