Task 3 : Multivariate regression

3.1 Carry out a multiple regression on all the independent variables, and determine the values for all the coefficients, and σ 2

```
Coefficients:
            Estimate Std. Error  t value Pr(>|t|)
(Intercept) -4.553658   5.830699   -0.781    0.435
X1          8.496442   0.005907 1438.265  < 2e-16 ***
X2          9.464434   1.108542    8.538  < 2e-16 ***
X3          9.370785   7.674613    1.221    0.223
X4          9.495708   0.930752   10.202  < 2e-16 ***
X5          9.110205   1.229118    7.412 5.44e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 26.96 on 494 degrees of freedom
Multiple R-squared:  0.9998,	Adjusted R-squared:  0.9998
F-statistic: 4.669e+05 on 5 and 494 DF,  p-value: < 2.2e-16
```

Value of a_0	-4.553658
Value of a_1	8.496442
Value of a_2	9.464434
Value of a_3	9.370785
Value of a_4	9.495708
Value of a_5	9.110205
Value of s^2	726.8416

3.2 Based on the p-values, R2 , F value, and correlation matrix Σ, identify which independent variables need to be left out (if any) and go back to step 3.1.

X	P	R^2	F
X1	2.2e-16	0.9997	1.76e+06
X2	0.4107	0.0004259	0.6779
X3	0.6453	0.0004259	0.2122
X4	0.3788	0.001556	0.776
X5	2.351e-14	0.1104	61.82

Correlation Matrix

	X1	X2	X3	X4	X5	Y
X1	1.00	-0.04	0.02	0.04	0.33	1.00
X2	-0.04	1.00	-0.01	-0.78	0.06	-0.04
X3	0.02	-0.01	1.00	-0.01	0.08	0.02
X4	0.04	-0.78	-0.01	1.00	-0.04	0.04
X5	0.33	0.06	0.08	-0.04	1.00	0.33
Y	1.00	-0.04	0.02	0.04	0.33	1.00

Based on the P values, R Square values, F values and the correlation matrix we can remove independent variables X2, X3 and X4 because :

1) There P values are greater than 0.05. This means the coefficient is equal to or close to zero , there is no effect. A low p-value (< 0.05) indicates that you can reject the null hypothesis. In other words, a predictor that has a low p-value is statistically significant as the changes in the predictor's value are related to the changes in the response variable and a larger p-value suggests that changes in the predictor are not associated with changes in the response and thus is statistically insignificant. Low P value means we fail to reject the null hypothesis(H_0).

2) There R Square values are quite low and close to zero . The coefficient of determination, or the coefficient of multiple determination for multiple regression needs to be higher for a good fit.

3) Correlation Matrix shows a strong negative correlation between X2 and X4 but they are bth discarded from the final model.

Now, our final model becomes:

$$Y = a0 + a1X1 + a2X5 + \varepsilon$$

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	44.540412	2.613336	17.044	< 2e-16	***
X1	8.496310	0.006476	1311.913	< 2e-16	***
X5	9.523692	1.341205	7.101	4.31e-12	***

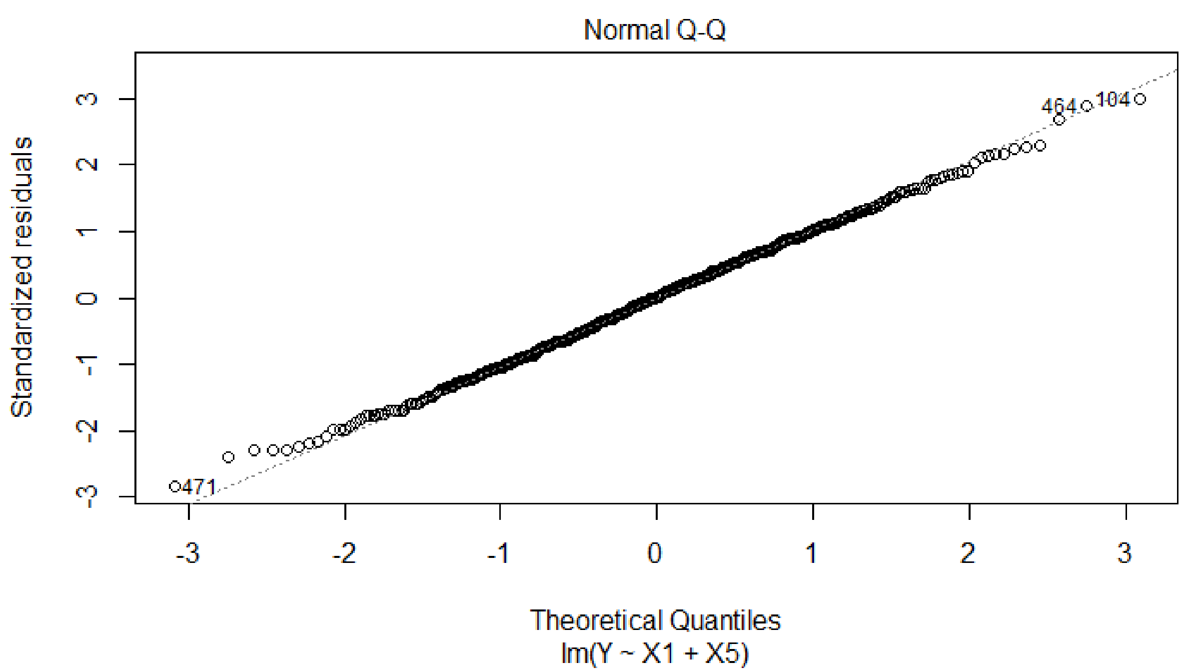
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 29.62 on 497 degrees of freedom
Multiple R-squared: 0.9997,	Adjusted R-squared: 0.9997
F-statistic: 9.674e+05 on 2 and 497 DF, p-value: < 2.2e-16

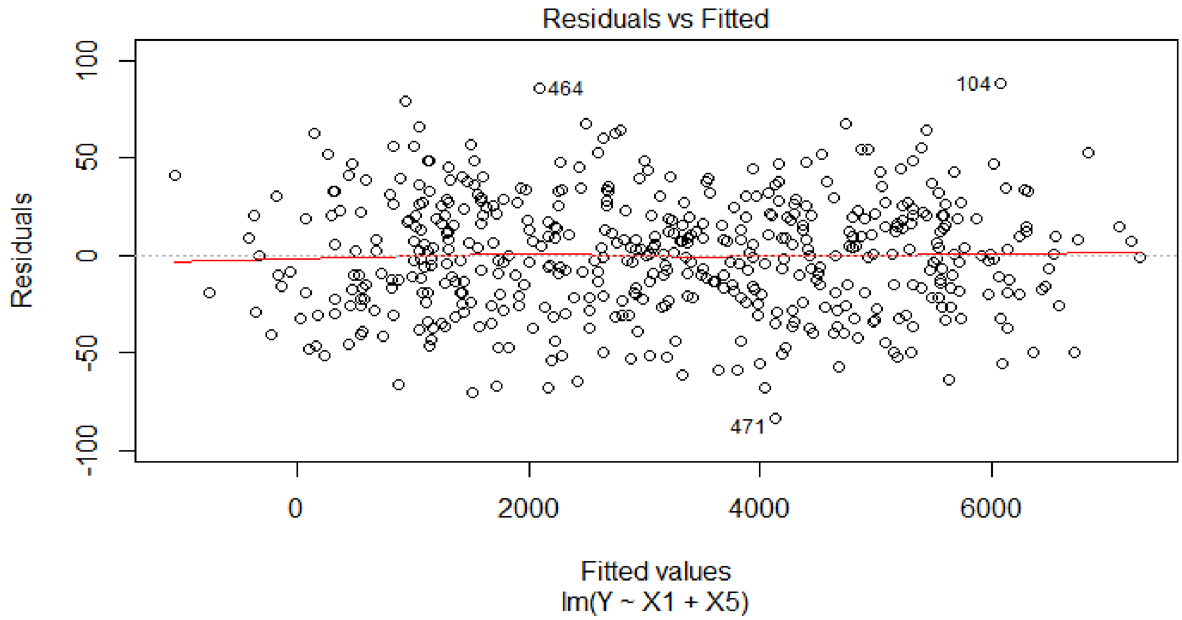
Value of a_0	44.540412
Value of a_1	8.496310
Value of a_2	9.523692
Value of s^2	877.3444

3.3 Do a residuals analysis: a. Do a Q-Q plot of the pdf of the residuals against N(0, s 2). Alternatively, draw the residuals histogram and carry out a χ 2 test that it follows the N(0, s 2). b. Do a scatter plot of the residuals to see if there are any correlation trends.

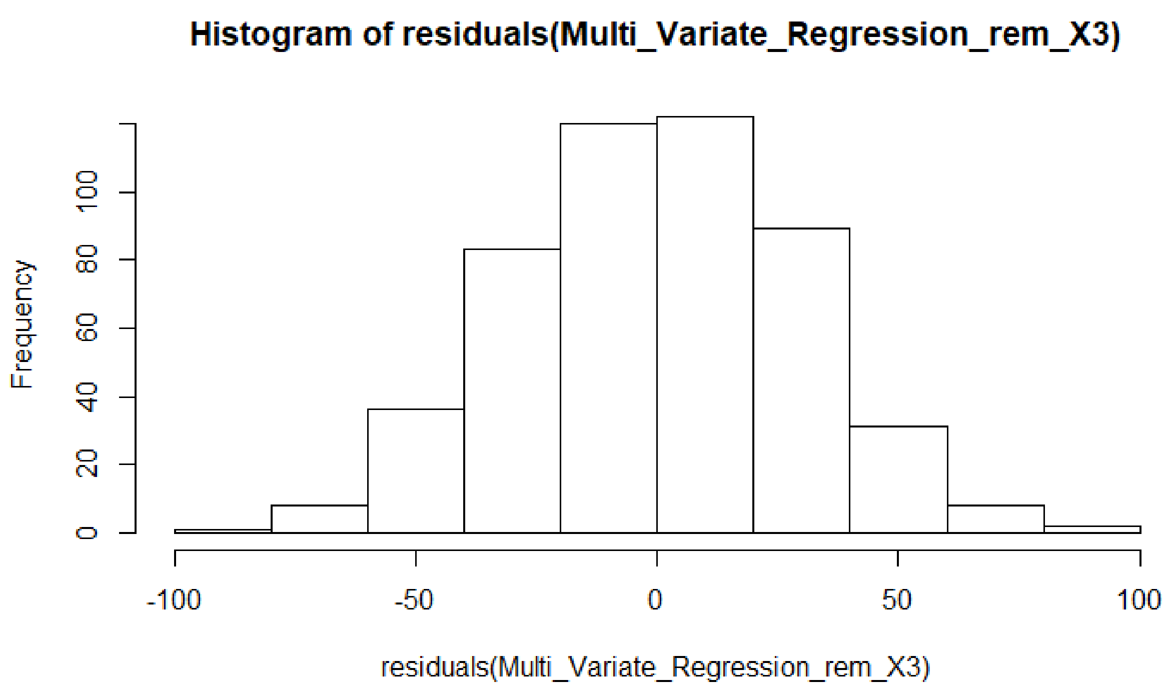
Final Model



As per the Q-Q plot, the residuals generated by the fitted values are normally distributed. They are a bot skewed near the ends but do not contain any outliers.



The above scatter plot shows that all residuals are centered around residual = 0 line throughout the range of fitted values. Thus, the residuals are not correlated and are random. This implies it is a good fit model. The residuals are normally distributed.



The histogram shows how the residuals are normally distributed.