Task 2: Linear regression

Before proceeding with the multiple regression, you will carry out a simple linear regression to estimate the parameters of the model: Y = a0 + a1X + ε, where X = X1.

2.1 Determine the values for a0, a1, and s 2 .

X1

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 42.632494   2.725424   15.64   <2e-16 ***
X1           8.511372   0.006416 1326.69   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 31.05 on 498 degrees of freedom
Multiple R-squared:  0.9997,            Adjusted R-squared:  0.9997
F-statistic: 1.76e+06 on 1 and 498 DF,  p-value: < 2.2e-16
```

Value of a_0	42.632494
Value of a_1	8.511372
Value of s^2	964.1025

2.2 Check the p-values, R2 , F value to determine if the regression coefficiants are meaningfull.

X	P	R^2	F
X1	2.2e-16	0.9997	1.76e+06

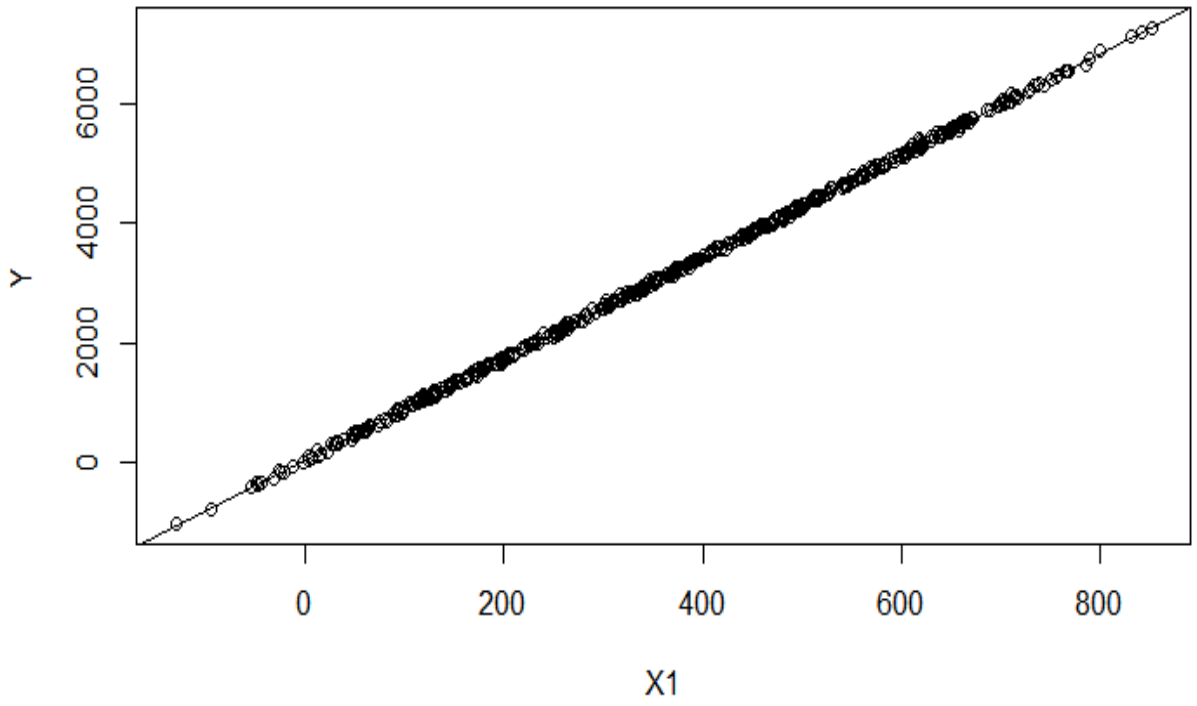
R-squared is a statistical measure of how close the data are to the fitted regression line. It is also known as the coefficient of determination, or the coefficient of multiple determination for multiple regression.
The higher the R-squared, the better the model fits your data. So, from the above table we can see that X1 is a good fit

The **p-value** tests the null hypothesis(H=0). If 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. In the above given table,

we can see that X1 has a P value that is quite low than 0.05, so we can consider it as statistically significant.

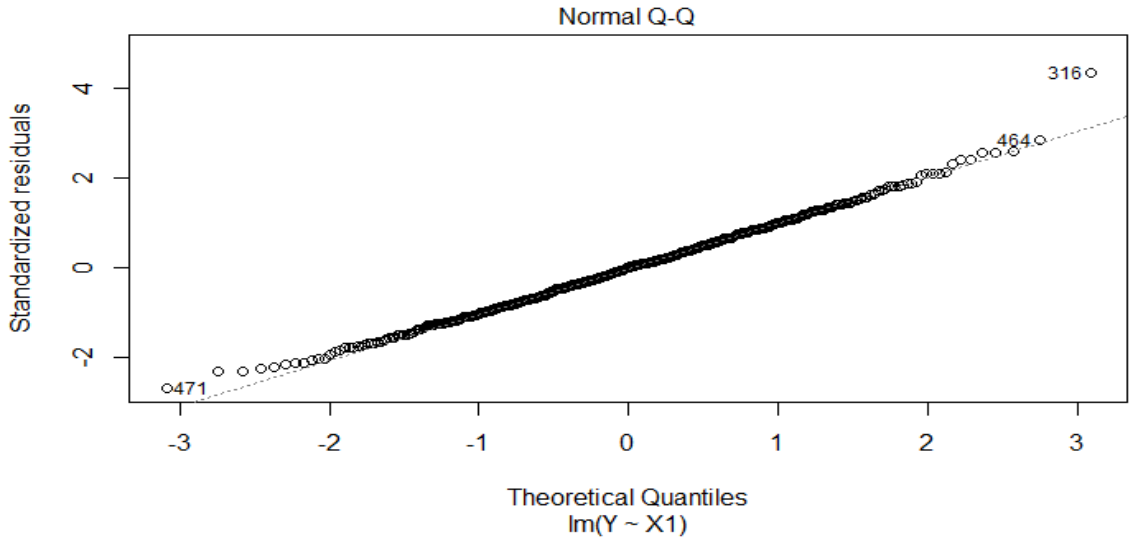
The **"F value"** statistics test the overall significance of the regression model. It checks if all the regression coefficients are zero. This tests the full model against a model with no variables and with the estimate of the dependent variable being the mean of the values of the dependent variable. The F value is the ratio of the mean regression sum of squares divided by the mean error sum of squares. Its value will range from zero to an arbitrarily large number. If F-Statistic value is significant, it gives extra confidence over R square values.

2.3 Plot the regression line against the data.

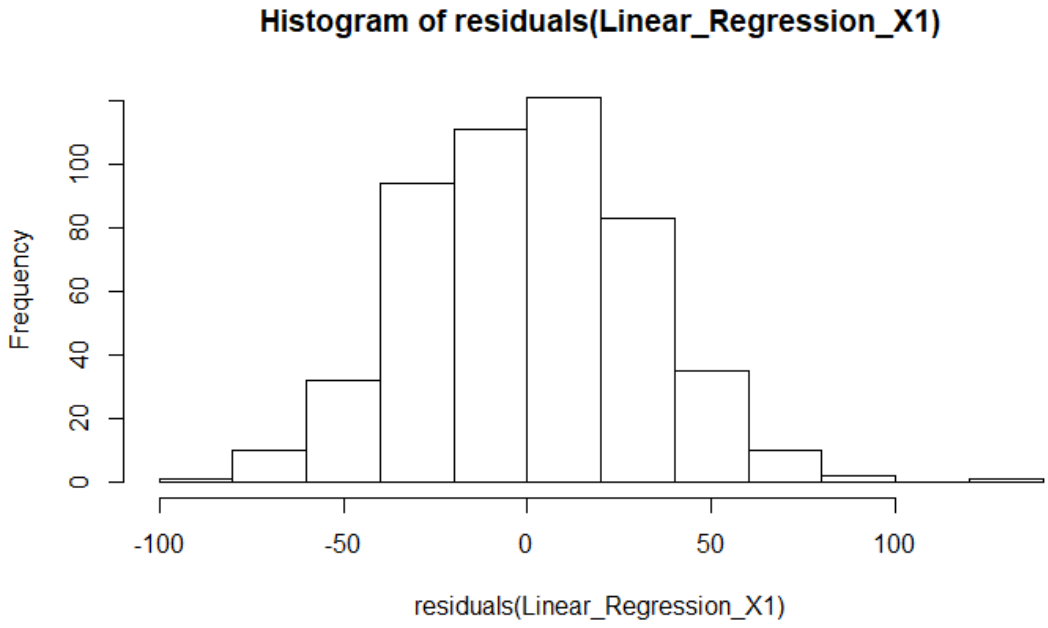


2.4 Do a residuals analysis:

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).

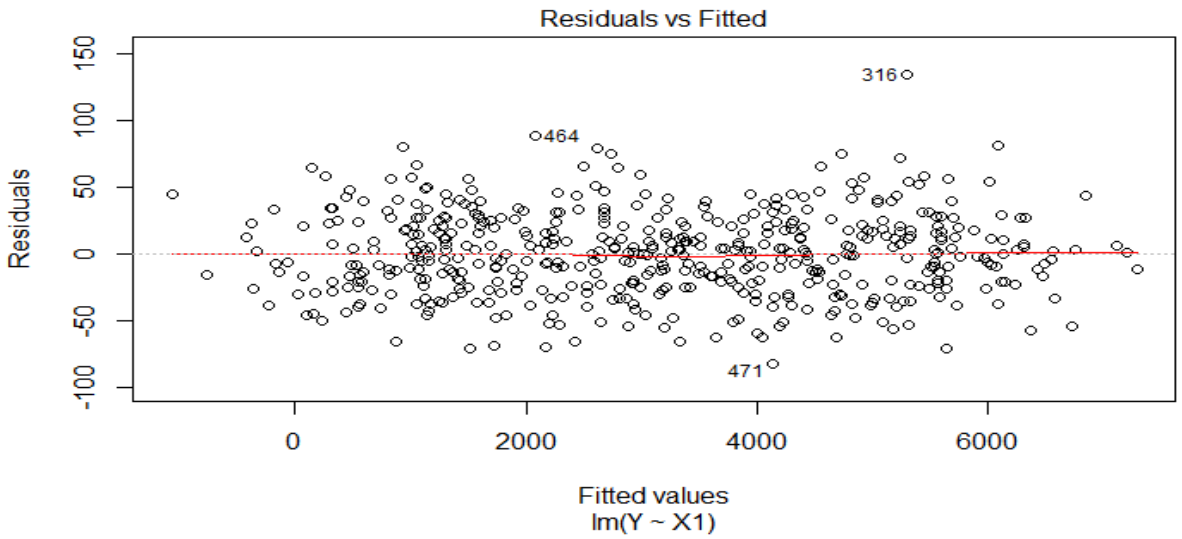


According to the Q-Q plot, the residuals are normally distributed and skewed at the ends. There are no outliers. This looks like a good fit model.



The histogram shows that the residuals are normally distributed.

b. Do a scatter plot of the residuals to see if there are any correlation trends.
The residuals should not be either systematically high or low. The residuals should be centered on zero throughout the range of fitted values. The model is correct on average for all fitted values and random errors are assumed to produce residuals that are normally distributed.
The plot is used to detect non-linearity, unequal error variances, and outliers.



For X1, residuals are centered around residual = 0 line.
Residual = observed – predicted
So, this meets our requirements and there is no correlation as we can't see in increasing or decreasing slope. Hence , linear model is apt for X1.

2.7 Use a higher-order polynomial regression, i.e., $Y = a0 + a1X + a2X^2 + \varepsilon$, to see if it gives better results.

For X=X1 ,
Linear Regression :

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	42.632494	2.725424	15.64	<2e-16	***
X1	8.511372	0.006416	1326.69	<2e-16	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 31.05 on 498 degrees of freedom

Multiple R-squared: 0.9997, Adjusted R-squared: 0.9997
F-statistic: 1.76e+06 on 1 and 498 DF, p-value: < 2.2e-16

Value of a_0	42.632494
Value of a_1	8.511372
Value of s^2	964.1025
Value of P	2.2e-16
Value of R^2	0.9997
Value of F	1.76e+06

Polynomial Regression :

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	4.414e+01	3.842e+00	11.490	<2e-16	***
X1	8.499e+00	2.309e-02	368.107	<2e-16	***
X1_Sq	1.669e-05	2.993e-05	0.558	0.577	

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 31.07 on 497 degrees of freedom
Multiple R-squared: 0.9997, Adjusted R-squared: 0.9997
F-statistic: 8.788e+05 on 2 and 497 DF, p-value: < 2.2e-16

Value of a_0	4.414e+01
Value of a_1	8.499e+00
Value of a_2	1.669e-05
Value of s^2	965.3449
Value of P	0.577
Value of R^2	0.9997
Value of F	8.788e+05

When we compare both models, P value for polynomial regression model is quite high, more than 0.05. So, this becomes statistically insignificant and we conclude that linear model is better. If we compare R- Square values, they are exactly the same .

2.8 Comment on your results in a couple of paragraphs.

Linear Regression is a linear approach to describe the relationship between a dependent variable and an independent variable. In this task we have 1 independent variable X1 and 1 dependent variable Y.

We performed linear regression , $Y= a0+a1X1+ \varepsilon$, where a0 is the intercept and ε is the error term and a1 is the coefficient of X1. The residual line is plotted against the data and the P values, R^2 values and F statistic values are calculated. These values were used to show how X1 is statistically significant and the coefficients are meaningful.

We got a0 and a1 and did the residual analysis. We found that the residuals are normally distributed for all the fitted values . We plotted the Q-Q plot, histogram of the residuals and the scatter plot to support our findings.

Then we also created a polynomial regression model over X1,
$Y = a0 + a1X1 + a2X1^2 + \varepsilon$
And did the regression and analysed the coefficient values. But the P value was quite large than 0.05. Hence we concluded that the linear model was better.