ASSIGNMENT 1 – DS 203

LINEAR REGRESSION AND ERROR ANALYSIS

Name : S Harikrishnan

Roll no : 24B2224

The objective of this assignment was to apply simple linear regression to a provided dataset and analyse the resulting model errors. The dataset contained two variables: *number of hours studied* (independent variable) and *exam marks* (dependent variable).

Two regression models were constructed:

1. **Standard Linear Regression** – Both slope and intercept were estimated from the data.
2. **Regression Through the Origin** – The model was constrained to pass through the origin (zero intercept).

For each model, a scatter plot of the given data points was created, with the corresponding regression line superimposed. The prediction errors for the standard linear regression model were further analysed by plotting a histogram of the errors to visualise their distribution. The skewness and kurtosis of the error distribution were calculated to assess its deviation from normality.

Part A

Given a set of observations which seem to have a linear relationship we would like to establish a regression line between x and y.

$$\hat{y}_i = \beta_0 + \beta_1 \cdot x_i$$

Where

$$\beta_0 = \frac{\overline{x^2}\bar{y} - \overline{xy}\cdot\bar{x}}{\overline{x^2} - \bar{x}^2}$$

$$\beta_1 = \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - \bar{x}^2}$$

Which are obtained by minimising the sum of squared errors.

Part B

1. The given dataset consists of 150 rows (observations) and 2 columns (variables) which are :

a) Hours studied:

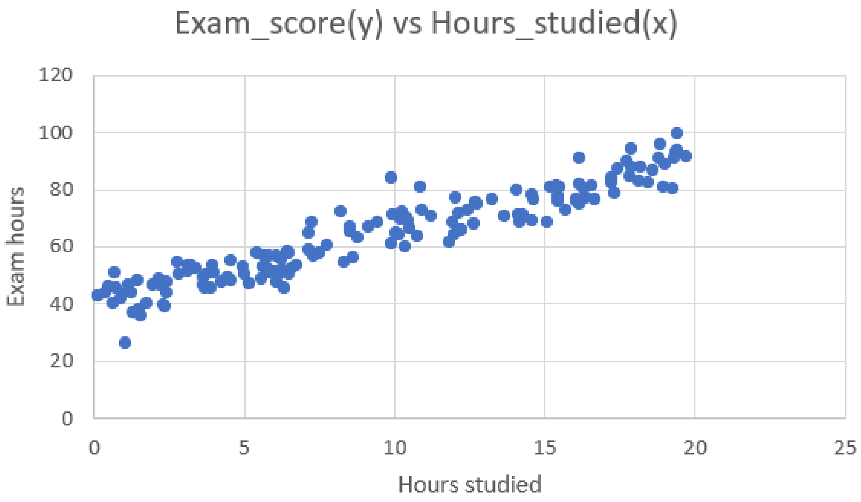
- It contains **numeric** type of data.

- Level of measurement: **Ratio**, since the hours studied can be measured with a meaningful zero and are continuous. The duration of 2 hours is twice that of 1 hour.
- The values range from 0.11 to 19.74 hours (found using the MIN and MAX functions in Excel).
- It represents the number of hours a student studied for the exam.

b) Exam score: **numeric**

- It also contains type of data.
- Level of measurement: **Ratio**, since exam scores have an absolute zero and are continuous.
- The values range from 26.25 to 99.5.
- It represents the marks the student scored in the exam.

2.



The scatter plot shows that as hours studied increases, the exam score tends to increase as well. The points are clustered around what appears to be a **straight-line relationship**, with not so many **outliers**. The relationship also looks to be approximately **linear**. So linear regression is a suitable model for this dataset.

3. In simple linear regression we try to fit the line:

$$\hat{y}_i = \beta_0 + \beta_1 \cdot x_i$$

to the given data where :

$$x_i = \text{number of hours studied by i}^{\text{th}} \text{ student}$$

$$y_i = \text{marks scored by i}^{\text{th}} \text{ student}$$

The values of β_0 and β_1 are given by :

$$\beta_0 = \frac{\overline{x^2}\bar{y} - \overline{xy}\cdot\bar{x}}{\overline{x^2} - \bar{x}^2}$$

$$\beta_1 = \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - \bar{x}^2}$$

where:

$\bar{x}$ = mean of $x_i's$

$$\bar{y} = \text{mean of } y_i\text{'s}$$
$$\overline{x^2} = \text{mean of } x_i^2\text{'s}$$
$$\overline{xy} = \text{mean of } x_i y_i$$
$$\bar{x}^2 = \text{square of } \bar{x}$$

These statistical parameters were calculated using built-in functions in excel and are given as:

Parameter	Value
$\bar{x}$	**9.457467**
$\bar{y}$	**63.87333**
$\overline{x^2}$	**124.3786**
$\overline{xy}$	**692.9502**
$\bar{x}^2$	**89.44368**

Substituting these values into the equation, we get :

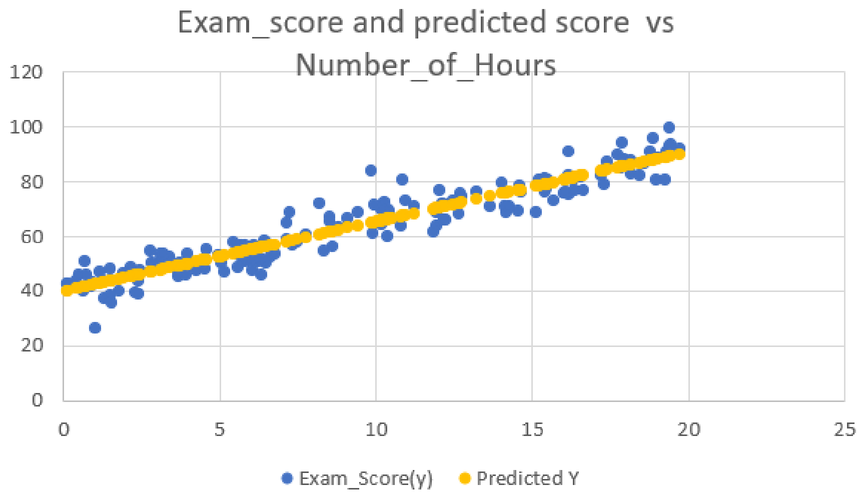
$$\beta_0 = 39.81467$$

$$\beta_1 = 2.543881$$

Hence, the equation of linear regression is :

$$\hat{y}_i = \mathbf{2.543881}x_i + \mathbf{39.81467}$$

4.



5.

6. **SSE**: The **Sum of Squared Errors(SSE)** quantifies the total squared difference between the actual and predicted values. Lower SSE indicates a better fit.

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)\hat{}2$$

Here, n = 150.

MSE: MSE stands for **Mean Squared Error.**

$$MSE = \frac{SSE}{n}$$

RMSE : It is the root of MSE (**Root Mean Squared Error**).

$$RMSE = \sqrt{MSE}$$

MAE : **Mean Absolute Error** (MAE) is the average of the absolute values of the differences between the actual and predicted values.

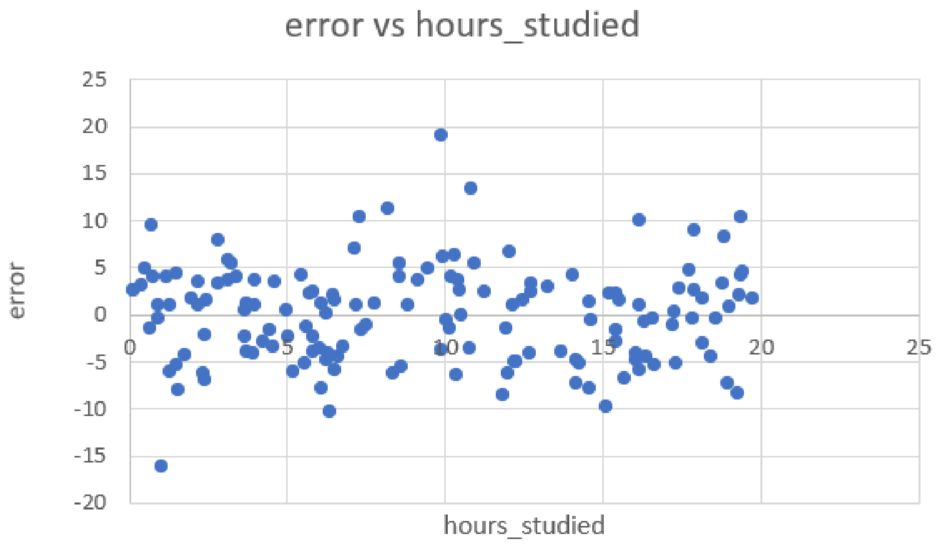
$$MAE = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y_i}\,|$$

It is **less sensitive** to outliers than SSE because it does not square the errors. It is not used for optimisation because since it involves the absolute value function which is **not differentiable** at x = 0.

The values of the parameters are given below :

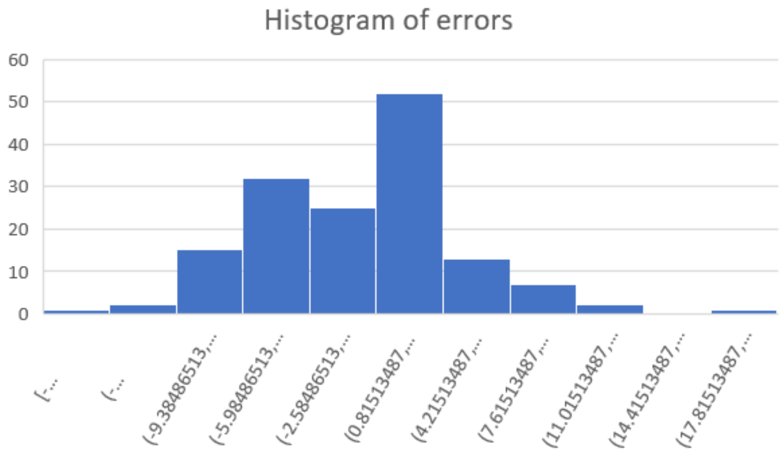
SSE	**3960.05**
MSE	**26.40034**
RMSE	**5.138126**
MAE	**4.112138**

7. Scatter plot of e_i vs x_i



The errors are **random** and **unpredictable** as the model has **extracted** out all the patterns in the dataset.

8.



Ideally, for a good regression, the histogram should :

- be **symmetric** and **centred around zero** : implies that the errors are **random**
- be **bell-shaped** (normal distribution) : A regression error (residual) is often the result of many tiny, independent influences on y that the model does not capture. So, by **Central Limit Theorem (CLT),** the errors should follow a normal distribution.
  The CLT states that :
  If you add up (or average) a large number of independent, small, random effects, their sum will tend to follow a Normal (bell-shaped) distribution, even if each effect is not normal individually.

The above histogram :

- appears roughly **bell shaped** with most of the errors centred around zero.
- does **not** have a huge **asymmetry**.
- looks approximately **normal** with a peak around **zero**.
- has **extreme values**

From the error analysis point of view, it is a **good** regression model.

9. **Skewness** and **kurtosis** are parameters that help us check whether a distribution is close to normal without directly fitting it.

Skewness: It is the **measure** of **asymmetry** in a distribution.
Skewness can be positive, negative or zero.

- **Positive** skew indicates that the **tail** on the **right** side of the distribution is **longer** than that on the left side.
- **Negative** Skewness means the **tail** of the **left** side of the distribution is **longer** than that on the right side
- **Zero** skewness indicates **symmetric** distribution. It signifies that the distribution of data is evenly distributed around the mean, with **no long tails** on either end of the distribution.

Since the **ideal normal distribution** is **perfectly symmetric**, its skewness should be zero. In practical cases, for normality **|skewness| should be close to 0**.

Kurtosis : It is a **measure** of how much the distribution's data points are **concentrated** around the **mean(peak)** or in the **tails (extreme values)** compared to a normal distribution. The kurtosis value for a normal distribution is 3. The type of kurtosis is determined by the value of **excess kurtosis** which is defined as :

Excess kurtosis = Kurtosis – 3

For a normal distribution, excess kurtosis = 0.

Kurtosis is mainly of three types:

- **Mesokurtic**: **Kurtosis ≈ 3** and **excess kurtosis ≈ 0**. It is shaped like a **normal distribution**.
- **Leptokurtic**: **Kurtosis > 3** and **excess kurtosis > 0**. It has **sharper peak** and **heavier tails** than normal distribution. More extreme values are likely.
- **Platykurtic**: **Kurtosis < 3** and **excess kurtosis < 0**. It has **flatter peaks** and **lighter tails** than normal distribution. Extreme values are less likely.

The skewness and excess kurtosis values of the error distribution are :

Skewness	**0.26146**
Excess kurtosis	**0.871417**

The **skewness** is **slightly positive** but **close to zero**, which implies that the distribution has a tiny **right tail**, but almost **symmetric**. This is evident from the histogram: most values are concentrated near the centre, while a few higher values extend the right tail slightly.

The **excess kurtosis** is **greater than zero**, which means it has a **sharp peak** and **heavy tails** which is also evident from the histogram where there are bins in the extremes on both the sides.

Conclusion : The distribution is approximately normal and symmetric with a tiny right tail, sharp peak and heavy tails.

10. **R²** is the notation for the **coefficient of determination** in statistics. It represents the **proportion of variation** in the dependent variable (y) that is accounted for by the regression line, compared to the **variation explained by the mean of y**. Essentially, it measures how accurately the regression model predicts for each independent variable compared to simply using the average value of dependent variable.

$$\boldsymbol{R^2 = 1 - \frac{SS_{res}}{SS_{tot}}}$$

Where SS_{res} = sum of the squares of the residuals (errors) = SSE

SS_{tot} = sum of the squares of the deviation of the values from the mean = total sum of squares

$$SS_{tot} = \sum_{i=1}^{n} (y_i - \bar{y})\text{^}2$$

For example, if R^2 = 0.85, it means 85% of the variation in the output is explained by the model, and the remaining 15% is due to factors the model doesn't capture.

An **ideal regression model** would have **R^2 close to 1** which implies that the model explains almost all the variability in the data. So, the value of R^2 for a regression model should be **as close to 1 as possible**.

The values of SS_{res} and SS_{tot} were found to be :

$$SS_{res} = 3960.05\ , SS_{tot} = 37871.34$$

The value of R^2 was calculated and found to be:

$$\mathbf{R^2 = 0.895434}$$

9. Sometimes, we are only interested in the slope of the regression model. For example, if we are not interested in the actual value of the variable and only want to know how much the dependent variable changes with the change in independent variable, calculating only the slope would also save computation.

For example, if we consider advertising expenditure vs sales revenue, marketing teams mainly want to know the effectiveness of advertising spend, that is how much additional sales each dollar of advertising brings.

a. Let the regression model be :

$$\hat{y}_i = \beta_1 \cdot x_i$$

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)\text{^}2$$

$$= \textstyle\sum_{i=1}^{n} (y_i - \beta_1 x_i)\text{^}2$$

Minimising SSE by taking derivative with respect to β_1 and equating it to 0.

$$\frac{d(SSE)}{d\beta_1} = 0$$

$$\textstyle\sum_{i=1}^{n} \frac{d(y_i^2 + \beta_1^2 x_i^2 - 2\beta_1 x_i y_i)}{d\beta_1} = 0$$

$$\textstyle\sum_{i=1}^{n} 2\beta_1 x_i^2 - 2x_i y_i = 0$$

$$2\beta_1 \textstyle\sum_{i=1}^{n} x_i^2 - 2\sum_{i=1}^{n} x_i y_i = 0$$

$$\beta_1 = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$$

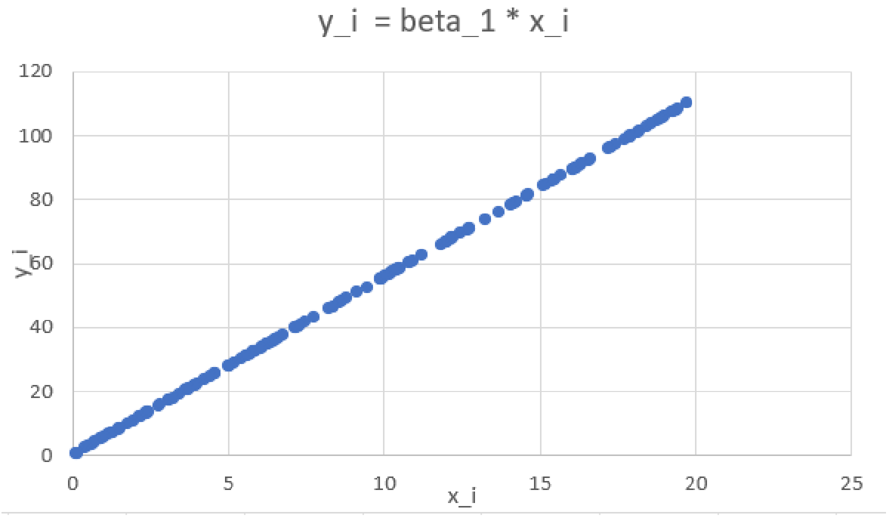
Substituting the values, we get

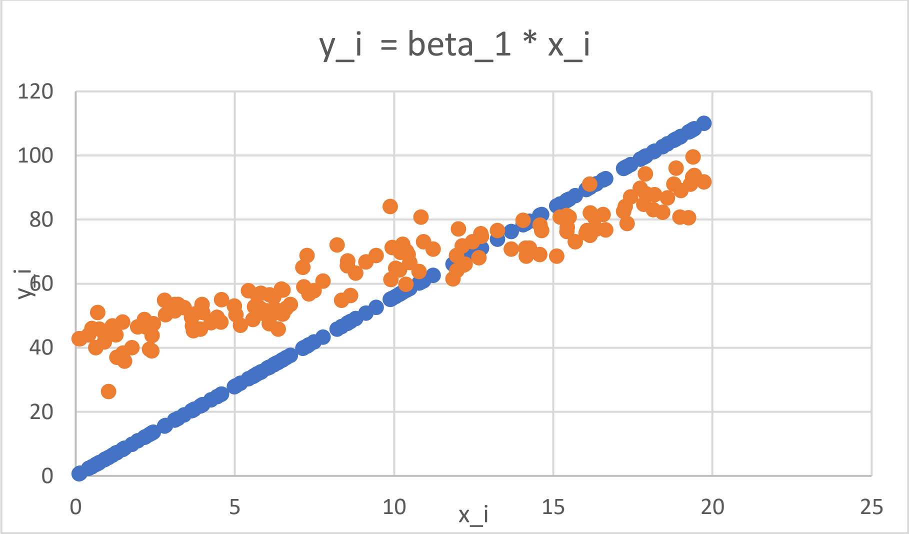
$$\boldsymbol{\beta_1} = \mathbf{5.571298}$$

So, the equation for regression model is :

$$y_i = 5.571298x_i$$

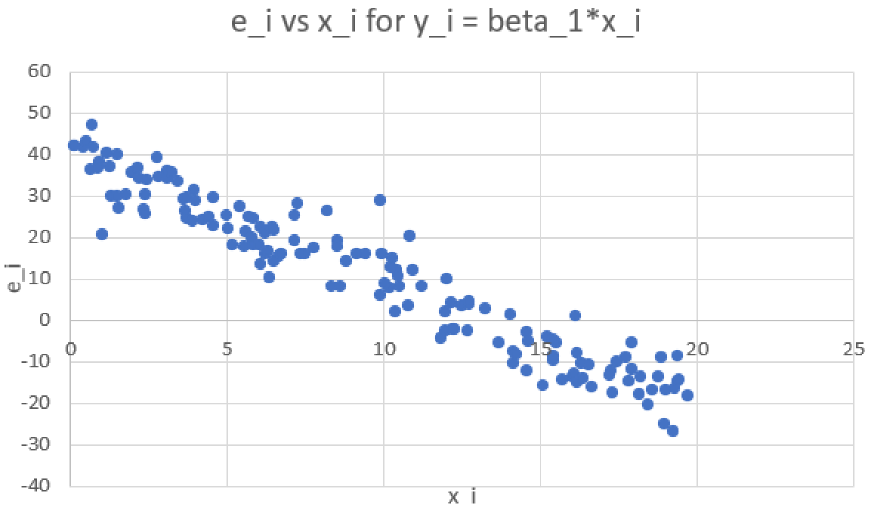
c.





Plot superimposed on the actual values

d.

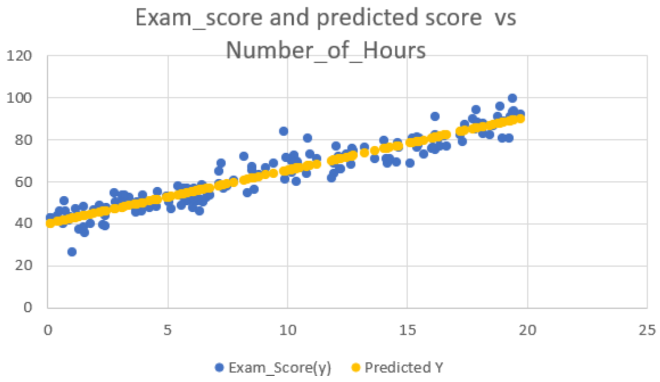


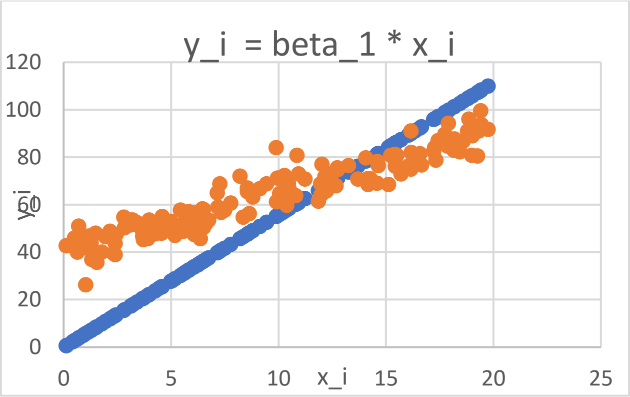
e. The values of the various parameters of the model $y_i = \beta_1 x_i$ have been listed below:

Parameter	Value
SSE	70747
MSE	471.647
RMSE	21.7174
MAE	18.6231

f. Comparison:

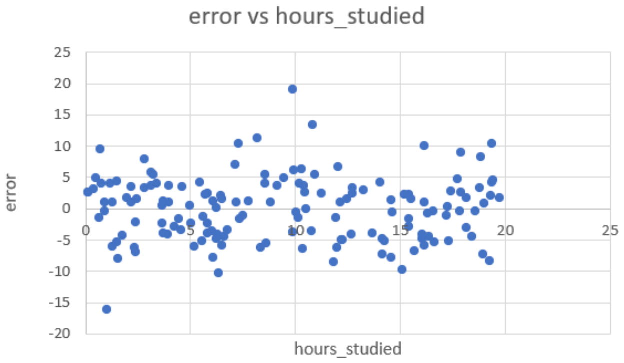
Plots:

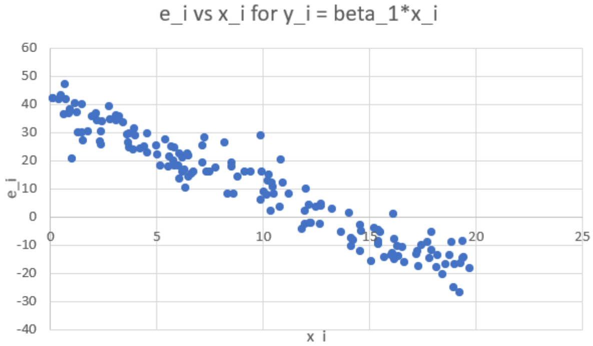




From the plot itself, it is very evident that the **first model** is a **better choice** compared to the second because the values predicted using the first model are closer to the actual values than those predicted by the second model . The values predicted by the second model are **distant** from the actual value especially when x_i is between 0 and 5.

Error plots:





A model is **good** if it has **extracted** out all the **patterns** in the data and the **residuals** are **random** and **unpredictable**. The error plot of the first model shows that the errors in the values predicted by that model do not follow any pattern and are random. The points are just scattered in the plot.

In the error plot of the second model, there is a very evident pattern which is similar to a straight line. This makes the error **predictable** and **not random**. Also, this means that the model has not extracted out the pattern completely from the data. Here we can at least predict that the error will be negative which is not possible from the error plot of the first model. Therefore, the first model is better than the second model.

Also, the **absolute values** of the **errors** in the values predicted by the **second model are greater** than that of the first .

SSE, MSE, RMSE :

SSE of first model = 3960.05

SSE of second model = 70747

SSE of first model < SSE of second model

The coefficients of the linear regression model are found so that the **SSE** is **minimum**. Still the SSE of second model is very large compared to that of first.

This implies that the errors in the first model are lesser compared to the errors in second model. The SSE model punishes the outliers which implies that the second model has a greater number of outliers than the first which is also evident from the plot.

Same is the case with MSE, RMSE since both are formed from the same dataset (n will be same).

MAE:
As evident from the plot, there are a greater number of outliers in the plot of the second model and than in the plot of the first model. So, the difference will be greater and hence the summation and average of the differences(which is the MAE) will be greater of the second model than that of the first.

A good regression model will have a low value for all these parameters and the errors in the values predicted by that model will be random.

Conclusion : The first model is better than the second model due to lower value of SSE, MSE, RMSE, MAE and also due to the random nature of the errors.

Main learnings:

- Learned to use excel to create linear regression model for the data
- Learned to analyse scatter plot
- Error analysis using the distribution of errors
- Learned about the different statistical parameters used to determine the normality of distribution such as skewness and kurtosis.
- Learned to compare two regression models using SSE, MSE, RMSE, MAE, R^2 scatter plots and the plots of errors, effect of intercept on error metrics