DS-203 Assignment
Autumn Semester 2025-26

ASSIGNMENT-01

SOURAV SREENATH
24B1841
B. Tech Engineering Physics
2^{nd} Year
Indian Institute of Technology, Bombay

Contents

1 Introduction

The given .csv file named E1.csv was analysed and regression lines were plotted. Mainly, two models were generated and they were compared to check which was the best one out of the two. This was done using few common statistical quantities which will be discussed later.

2 Model 1: Regr. line best fitting the plot

2.1 General Idea on the Data provided

The data provided contains two columns namely, Hours_Studied and Exam_Score. There are 150 data points for each quantity. This seems like a set of data points which compares the effort put in by students and the result, ie: marks or grades they got. The sample size is small enough for easy usage and handling in excel itself.

The levels of measurement that could be associated for both Hours_Studied and Exam_Score is Ratio. This is because both are continuous values itself and a proper zero point is defined for both. Comparison using ratios also make sense here. Example: Twice the number of hours, twice the marks, etc.

2.2 How was it done?

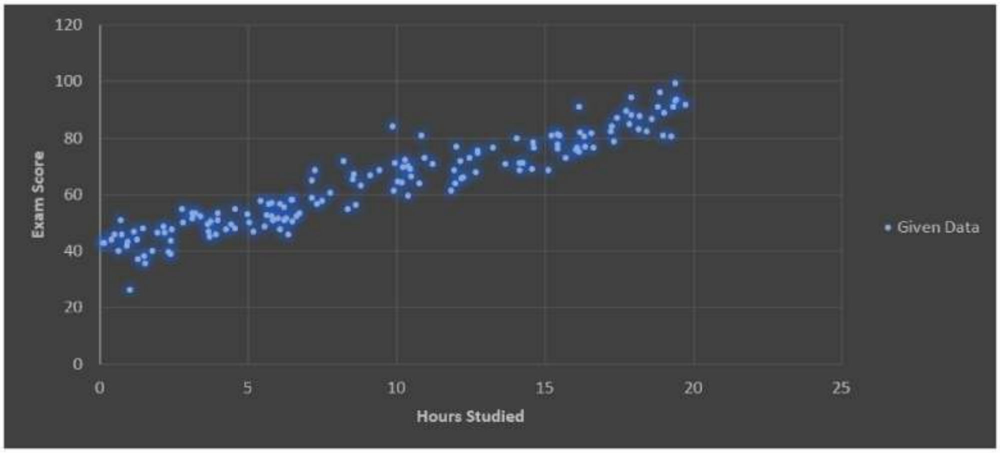


Figure 1: Scatter Plot of the Data

The data points, on a visual inspection appear to form a line. This means a model using Simple Linear Regression would work well and it could be done without using polynomials and risking overfitting and capture of noise. This is due to the bias-variance tradeoff which I had mentioned in my Google-Form analysis. It means that too little complexity means you don't capture enough of the pattern while too much complexity means our model would work exceptionally well with the given data, but due to noise uptake due to higher order, the prediction capabilities would be lost. Not slipping away from the topic too much, I have used SLR.

This was done using the formula that was discussed in class itself. The line was represented in slope intercept form and the parameters, ie: slope and intercept were calculated as per the formula. The formula is given below,

$$\beta_1 = \frac{\overline{xy} - \overline{x}\,\overline{y}}{\overline{x^2} - (\overline{x})^2} \tag{1}$$

$$\beta_0 = \frac{\overline{x^2}\,\overline{y} - \overline{x}\,\overline{xy}}{\overline{x^2} - (\overline{x})^2} \tag{2}$$

2.3 Finding the Regression Line

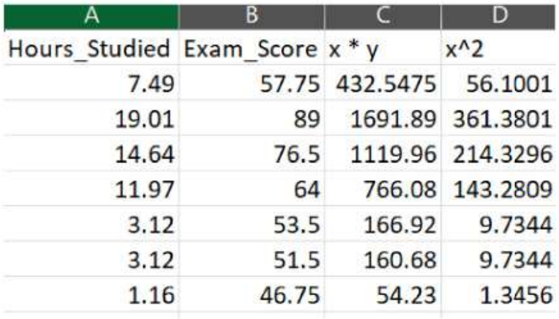
We know formulae for the slope and intercept of the regression line for any set of 2 data quantities. The formula is in the previous section.

To compute the value, I need the mean of xy, x^2, etc. Assuming x as the Hours_Studied and y as the Exam_Score, I made new columns labelled xy and x^2. I populated the values of these using the formula options in excel.

One learning opportunity came here. I paused my work in between and saved my file as a .csv file. But .csv files only store raw data, hence the formula i used to populate the cells is not visible for those cells I had already filled before saving. Only the numbers are visible, when I click the cells. Also, .csv files also don't save the plots generated. So, for the end submission, I will submit a .xlsx file to show the formula usage in the columns after the 'learning opportunity' and the figures.

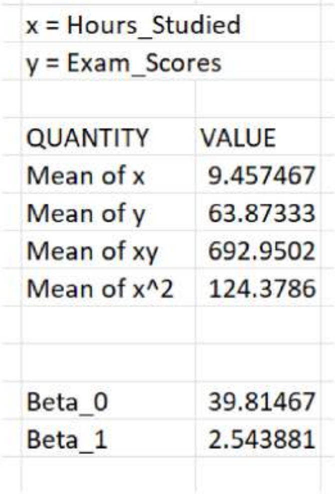
After getting the columns populated, I used a cell and ran the AVERAGE function in excel for those columns from Row2:Row151. Then I had the average of x, y, x^2 and xy.

Now I used another cell and used a formula for calculating β_0 and β_1 using the derived formula and the values of averages. This gave me the quantitative idea on how the regression line looks.



A	B	C	D
Hours_Studied	Exam_Score	x * y	x^2
7.49	57.75	432.5475	56.1001
19.01	89	1691.89	361.3801
14.64	76.5	1119.96	214.3296
11.97	64	766.08	143.2809
3.12	53.5	166.92	9.7344
3.12	51.5	160.68	9.7344
1.16	46.75	54.23	1.3456

Figure 2: Columns



x = Hours_Studied

y = Exam_Scores

QUANTITY	VALUE
Mean of x	9.457467
Mean of y	63.87333
Mean of xy	692.9502
Mean of x^2	124.3786
Beta_0	39.81467
Beta_1	2.543881

Figure 3: Averages

2.4 Plotting

I had two major ways to plot the line using the found out values of β_0 and β_1. One assignment anyways was to plots the regression line predicted y_i values and actual y_i values vs x_i.

Using the above fact, I could have changed the predicted scatter to a scatter plot with a line to get a continuous line. Or I could have added a new data set using the min and max x and y values as the x and y series. Both approaches I did try and both, ofcourse gave the same answer, but the former method is the easiest. I didn't use the already present trendline feature though to plot the graph but I did use it to check if my line was correct and indeed it was.

The two versions of plots are shown below,

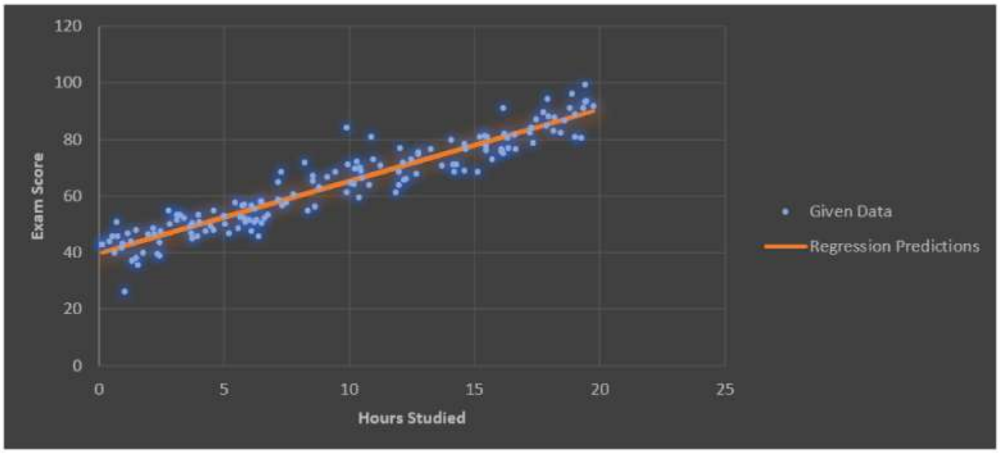


Figure 4: Scatter Plot with Reg. Line

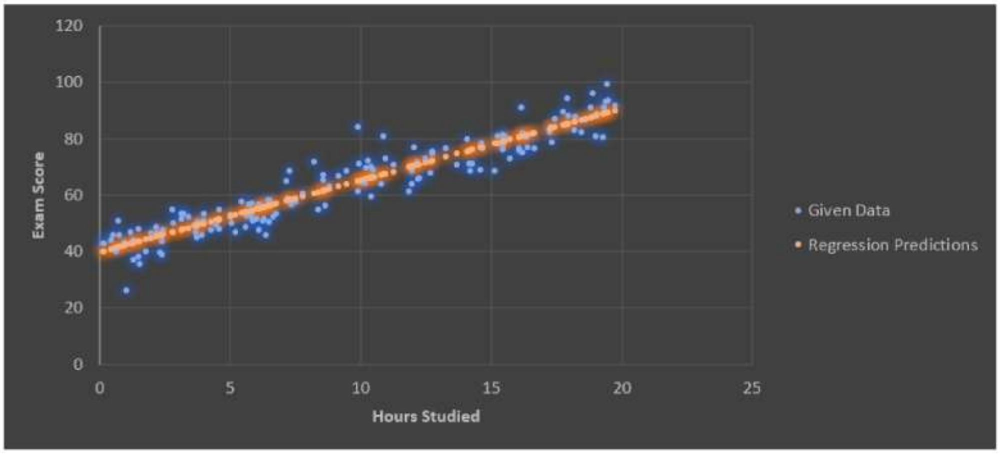


Figure 5: Scatter Plot with Reg. Line Predicted points

2.5 Error Analysis

The model has been created now, but to understand whether it is a good model or not, we have to see the residuals, ie: the errors which are left out. These give us a good idea on how well the model has squeezed out information out of the model. If most of the pattern has been made out by the regression model, then the error distributions look random and form a bell curve when looking at error only. If some pattern still appears for the error graph, then, more information is still left out and not taken up by the model. In a way the more unpredictable the residues, the more predictable is the future for our model.

So, error has been defined as the difference of the actual y value and the predicted y value, ie:

$$e_i = y_i - \hat{y}_i \tag{3}$$

here, y_i is the known y values for each x_i and $\hat{y}_i$ is the predicted y values for the x_i values.

The error can be positive and negative and analysing the errors with some statistical methods gives us a good idea on how well a model performs quantitatively.

Some of the terms that have been calculated by me are: SSE, MSE, RMSE and MAE. These have been calculated by forming new columns for squared error, absolute error, etc and putting formulae in excel to get their averages, sums, etc.

The formulae for the calculation of each value is given below,

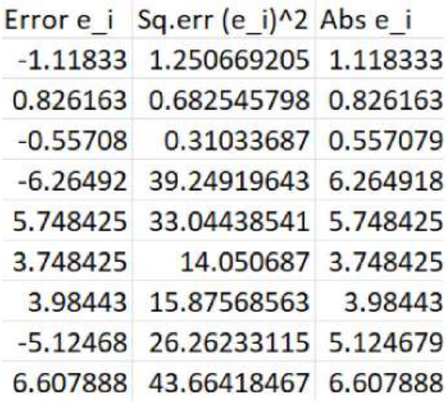
$$\text{SSE} = \sum_{i=1}^{n} (e_i)^2 \tag{4}$$

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} (e_i)^2 \tag{5}$$

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (e_i)^2} \tag{6}$$

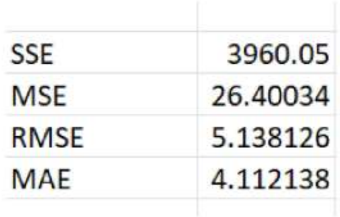
$$\text{MAE} = \frac{1}{n} \sum_{i=1}^{n} |e_i| \tag{7}$$

The calculation has been done by excel formulae and the answer is given on excel. Below is a screenshot of the answers I had got,



Error e_i	Sq.err (e_i)^2	Abs e_i
-1.11833	1.250669205	1.118333
0.826163	0.682545798	0.826163
-0.55708	0.31033687	0.557079
-6.26492	39.24919643	6.264918
5.748425	33.04438541	5.748425
3.748425	14.050687	3.748425
3.98443	15.87568563	3.98443
-5.12468	26.26233115	5.124679
6.607888	43.66418467	6.607888

Figure 6: Error related quantities calculated



SSE	3960.05
MSE	26.40034
RMSE	5.138126
MAE	4.112138

Figure 7: Quantities Calculated

All of the above quantities are good for overall usage but some terms are preferred more than other in some scenarios. For example,

- SSE is mainly used in model fitting according to one particular dataset. Optimization of this value guarantees that most points almost agree with the regression model. Smaller is the SSE, more is the closeness of all points as a whole to the line(in this case). SSE also helps in finding out other important terms like R^2.

- MSE is also used for model training but the key feature with MSE is that it is independent of scale. SSE was dependent on the number of data points in the sample whereas MSE is not. Where SSE was good at training models with one dataset, MSE can be use to train models with multiple datasets because the average of the SSE means that, there is no scaling or dependence on the number of data points.

- RMSE also could be used for training models, but RMSE has the privilege of having the same units as error itself. This means it can be it can be used in places where a lot of interaction based on such quantities take place, where if MSE was used, it would have made it much more complicated to work on two different units. An example could be weather predictions where RMSE is more important.

- MAE is used in places where a lot of outliers are present and use of SSE could absolutely result in huge values. Another interesting fact about the MAE is that, while SSE, MSE and RMSE help the line converge towards the mean value, the MAE helps the line converge towards the median. Line scenario used just as an example. So, if we want the median instead of the mean, MAE is the best error analysis technique.

2.6 Plots of Errors

To get a visual idea on how well the model works, we have to check the residual. If the error is unpredictable and normal, it means our model is quite good. For this, I used the built in Excel tools for histogram and scatter plot and added the x and y series as the columns containing e_i and x_i. The below is what I got,

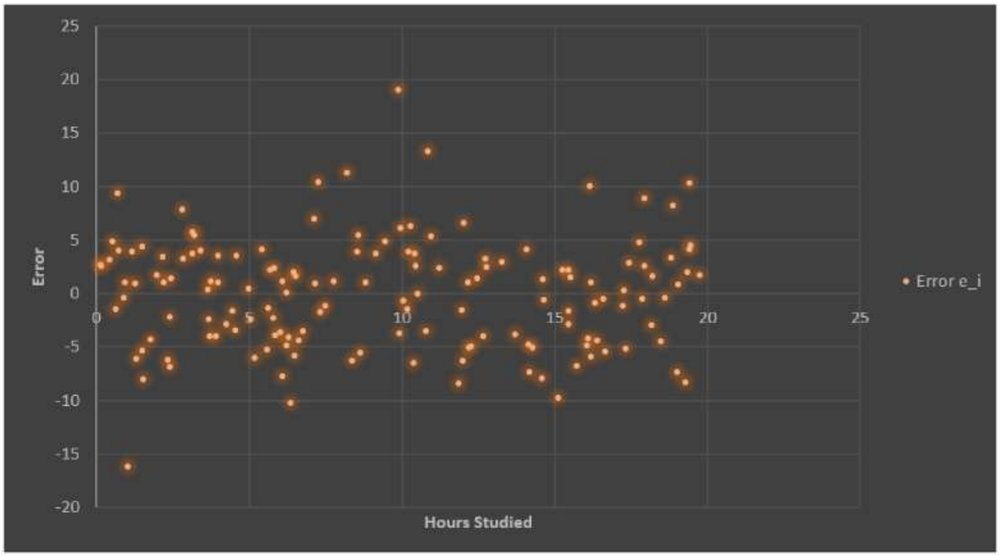


Figure 8: Error Scatter Plot

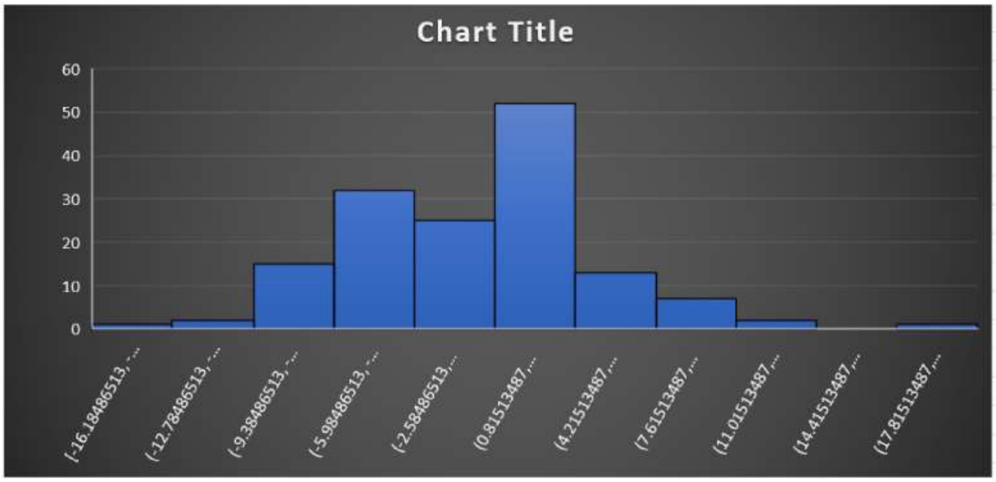


Figure 9: Error Histogram

Visual inspection reveals that model seems to be a good fit to the given data sets. This is because the residuals, ie: errors are quite unpredictable. I cannot see any exact pattern for the error scatter plot and therefore it visually appeals to me.

The histogram too, on visual inspection reveals an almost made bell curve shape but which seems to have some bimodal character to it. Even with that, the error values keep dropping out on either side. This means that the errors do follow a normal distribution pattern. So, this also seems fine.

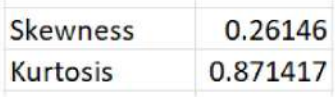
Visual inspections seems as though this model is quite good with the regression it is following.

Both the above statements were made very qualitatively. To get a quantitative and rigid idea onto whether this is possible, we must bring in terms like Skewness and Kurtosis too.

2.7 How well the model performs?

Skewness and Kurtosis are the two terms that would help us see whether the above model is a good model or not. They tell us how unpredictable the residuals are. Visually speaking, they check if the graph resembles a bell curve or if the mean is at 0, etc.

This calculation was directly by the built in excel function that is SKEW and KURT which measure Skewness and Kurtosis. I entered in the error e_i column values and put it into the function, which gave the below result.



Skewness	0.26146
Kurtosis	0.871417

Figure 10: Skewness and Kurtosis

These two values will help us quantitatively tell how good the model is.

Skewness refers to the asymmetry of the error histogram. This means that, it measures whether one side of the regression line contains more points than the other. Here, a skewness of 0.26 is measured which is very small and is fine for a model like this. This is fine and it means that the mean value of the errors is close to 0 itself with very little deviation. We want the curve to be as close to a normal distribution centred at 0 as possible.

Kurtosis has a value 0.87. Kurtosis gives a measure of the curve itself. In the sense, how fast values drop off, etc. In this case, a Kurtosis of 0.87 means that there is sharp concentrated error bar near the middle and any concentration anywhere on the tails too. This means the curve looks like its flattened a bit. This is why Kurtosis is 0.87 here. This is fine for a model as seen above. A very good reason why such a thing was seen might be because of the small sample size we are working with. Considering our sample size, such a Kurtosis value is not bad.

Thus, the Skewness and Kurtosis too reveal that the model is a good model.

Another good test to see how well the model performs is calculating the R^2 value. R^2 is a statistical quantity that measures the overall error and compares it to the overall variance of the dataset itself. It is a good measure to check how good a model is. Better R^2 values mean the model has lesser error values compared to the variance contained in the model itself. So, the thumb rule is, higher the value of R^2, better the model. Though, the thumb rule is this, it is always best to avoid too high R^2 values, as that might be an indication of overfitting and the model would work exceptionally well with the given data but poorly in predictions.

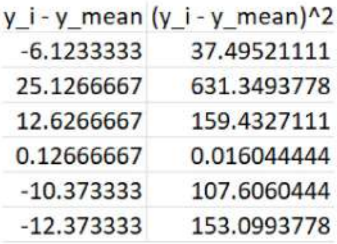
Calculation of R^2 value is done by using the below formula,

$$R^2 = 1 - \frac{\text{SSE}}{\text{SST}} \tag{8}$$

I calculated SST for the dataset which is,

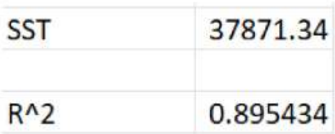
$$\text{SST} = \sum_{i=1}^{n} (y_i - \bar{y})^2 \tag{9}$$

An important note here is that, the SST is independent of the model always and only depends on the given sample itself. It pretty trivial once we get to know that it measures the variance of a particular sample.



y_i - y_mean	(y_i - y_mean)^2
-6.1233333	37.49521111
25.1266667	631.3493778
12.6266667	159.4327111
0.12666667	0.016044444
-10.373333	107.6060444
-12.373333	153.0993778

Figure 11: Columns used for SST calculation



SST	37871.34
R^2	0.895434

Figure 12: R^2 value

The R^2 value I got was close to 0.895. This is very close to 1 for practical purposes and not too close to 1 too, so no overfitting. This is because the sum of errors was small compared to the variance of the sample itself. (comparing to variance. Its not exactly variance though, but its closely related). This meant the model was pretty accurately able to predict each point of the data taking into account the internal variance the data had. This high value of R^2 also confirms that the model is a very good model for the sample given.

3 Model 2: Regr. line through the origin

3.1 Why is it done?

If we have some sets of data points, also assume a line is the best model, and when the regression is conducted, we see that the line may not go through the origin. Then in some cases, if we take another line from the origin, won't it actually give a worser model? Well, it is true that that could happen. But sometimes it is a necessity. Example, we are conducting an Ohm's Law experiment. Here, V is proportional to I. But our data points may not necessarily agree with this. But this proportionality is not a choice, it is a forced constraint. In such conditions, even if the model gets worse, we still line up the line with the origin. Another example could be Salary vs no.of hours worked.

Another situation where such a model would be taken is when we can't afford to take in cases with intercepts. This could be the case for complex calculations happen with the model. In such a case, a simple proportionality, even though could be a bit worse prediction wise, still could go a long way with managability.

3.2 Selection of the best line

Now we need to find the best line passing through the origin and minimizes the squared errors of the data. This can be done writing the line's equation as follows.

$$\hat{y}_i = \beta_1 x_i \tag{10}$$

Now we need to minimize the SSE to find the optimized solution. First we write the error e_i for that,

$$e_i = y_i - \hat{y}_i = y_i - \beta_1 x_i \tag{11}$$

Now the SSE,

$$SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \beta_1 x_i)^2 \tag{12}$$

Differentiating wrt β_1 to get the minimum,

$$\frac{d}{d\beta_1} SSE = \frac{d}{d\beta_1} \sum_{i=1}^{n} (y_i - \beta_1 x_i)^2 = \sum_{i=1}^{n} 2(y_i - \beta_1 x_i)(-x_i) \tag{13}$$

$$\sum_{i=1}^{n} 2(y_i - \beta_1 x_i)(-x_i) = 0 \tag{14}$$

$$\sum_{i=1}^{n} (y_i - \beta_1 x_i)(-x_i) = 0 \tag{15}$$

$$\sum_{i=1}^{n} -x_i y_i + \beta_1 \sum_{i=1}^{n} x_i^2 = 0 \tag{16}$$

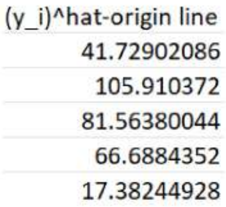
$$\beta_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i \tag{17}$$

The final equation for β_1 thus becomes,

$$\boxed{\beta_1 = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}} \tag{18}$$

3.3 Plotting the Regression Line

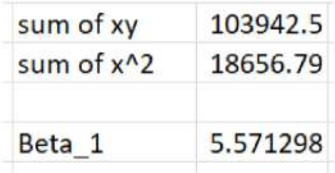
I calculated the value of β_1 the line would take with excel itself. I could either use the SUM function the columns or just multiply the already calculated averages by 150, ie: the number of data points.



(y_i)^hat-origin line
41.72902086
105.910372
81.56380044
66.6884352
17.38244928

Figure 13: New Columns

Then using the formulae in excel I found the below details,



sum of xy	103942.5
sum of x^2	18656.79
Beta_1	5.571298

Figure 14: Values

Thus, I got the required β_1 value and i plotted that like the last model and received the below two figures. Both the figures are the same with the only difference that one contains simple predicted points for each x_i whereas the other contains a continuous line itself.

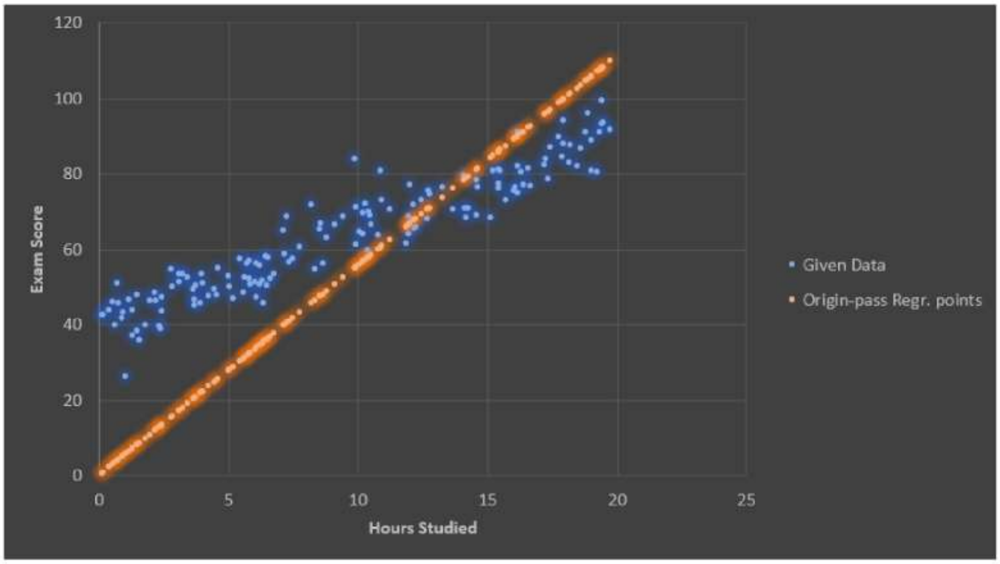


Figure 15: Points Regression Figure

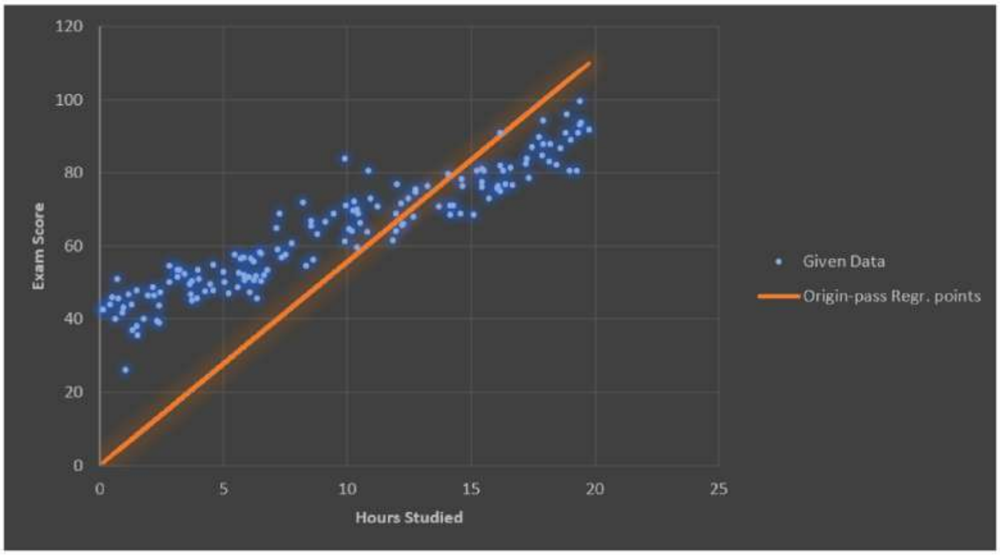


Figure 16: Line Regression Figure

3.4 Error Analysis

Similar to the last model, a look into the error scatter plot vs x_i and the error histogram reveal the following figures,

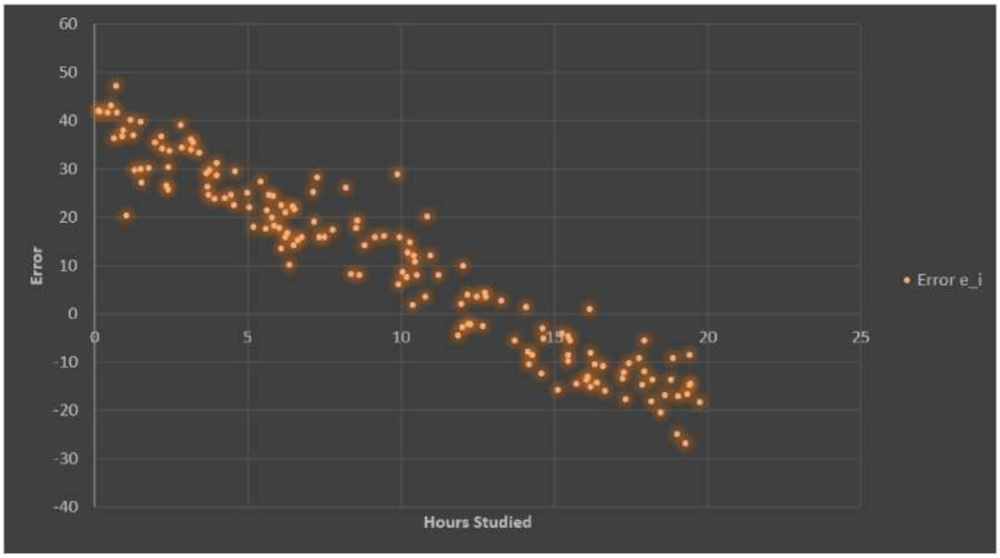


Figure 17: Error Scatter Plot

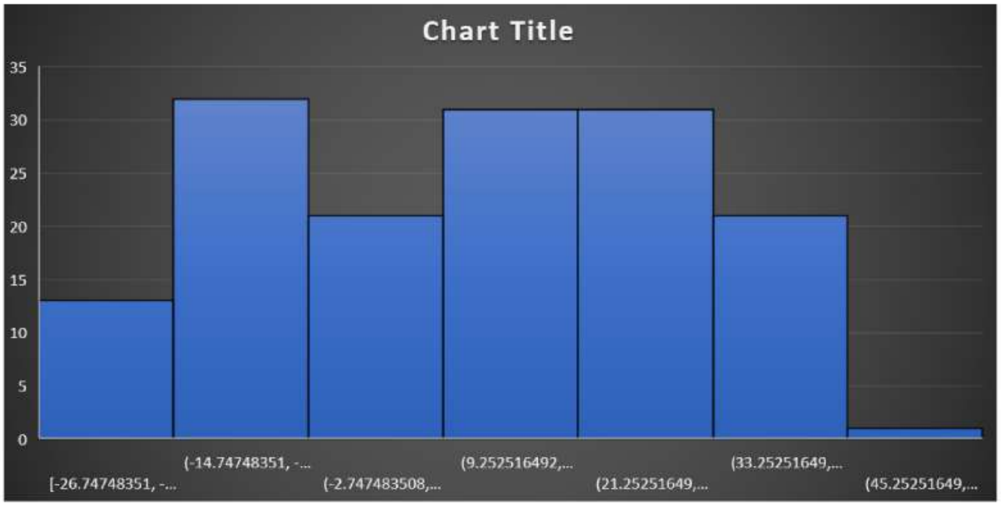
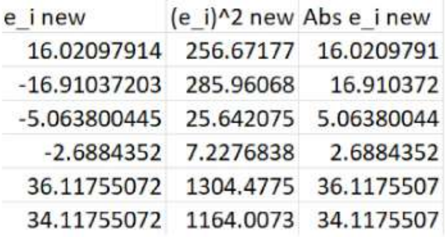


Figure 18: Error Histogram

Figuring out whether this model is a good model or not can be done in two ways visually or by math. First lets do the visual aspect. The error scatter plot here is highly predictable. In the sense, the residuals are not random, it has a clear pattern, ie: they form a line. This means the model has not squeezed out all the patterns out from the data provided. This is why the error has such a distinctive pattern. When the error become predictable, it means the actual model prediction become unreliable. The histogram too, there is no real peak, it seems almost constant throughout with few sinks, and the error reaches a maximum of more than absolute values of 45, etc, which looks very bad. It seems though that the model is performing terribly and fits the data very badly.

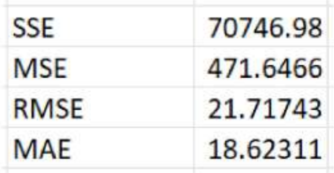
Now using mathematical tools,



e_i new	(e_i)^2 new	Abs e_i new
16.02097914	256.67177	16.0209791
-16.91037203	285.96068	16.910372
-5.063800445	25.642075	5.06380044
-2.6884352	7.2276838	2.6884352
36.11755072	1304.4775	36.1175507
34.11755072	1164.0073	34.1175507

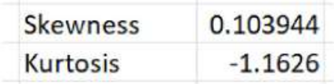
Figure 19: New Columns Used for Error Analysis

Applying excel formulae on these data, I got the following,



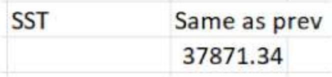
SSE	70746.98
MSE	471.6466
RMSE	21.71743
MAE	18.62311

Figure 20: SSE, MSE, RMSE, MAE



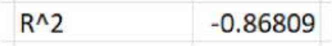
Skewness	0.103944
Kurtosis	-1.1626

Figure 21: Skewness and Kurtosis



SST	Same as prev
	37871.34

Figure 22: SST



R^2	-0.86809

Figure 23: R^2

Comparison in the next section.

The skewness seems fine while the Kurtosis has exceeded the magnitude we saw for our previous model. Comparison on this later. The highlight here is the R^2 term which has taken almost the same value as our previous model but in negative terms.

4 Comparison of the Models

Now, we have perceived how both the models look like and the values each one took when analysed by various mathematical tools.

Visually speaking the second regression line is quite far away from the actual trend whereas the first line lies inside the trend itself. This fact is quite obvious. The error scatter plot reveals that the errors in the second model are immensely huge compared to the first model. The histogram in the first case resembles a normal curve in some way even though bimodal. But, the second one seems completely off and seems almost flat with no clear maxima, with all being at almost the same level. The second model's residuals are highly predictable and hence the second model has not completely extracted the pattern out of the data.

Numerically, the intercept model had $R^2(= 0.895)$, SSE (=3960), RMSE (=5.14), and MAE (=4.11).

The origin model had R^2 (=-0.868) and SSE (=70746.98), which means it's actually worse than just using the mean of y as a prediction.

The slope from the intercept model fits the main cloud of points in the scatterplot quite well, but in the origin model the slope (=5.57) seems tilted in a way that mainly pleases the extreme x-values and misses the rest. Its seems as though the extreme x values have much more representation in this model, which i will come to later.

From the residual histograms, the intercept model's errors form a central peak near zero, while the origin model's histogram is much flatter. I learned that the intercept model's kurtosis (= 0.87) means it has a common error size close to zero, but the origin model's negative kurtosis (-1.16) means there's no single dominant error size, the model is more or less equally wrong for almost everyone.

The skewness values (0.26 vs 0.10) made me think the errors were balanced in both cases, but when I looked at the residual scatterplots, the origin model's residuals clearly sloped from positive at low x to negative at high x.

The reason why this actually occurs is because the errors are almost equal on either side of the number line. So, in a sense, the skewness didn't pick up any issue because the errors cancelled each other out so that the net result was close to 0. I researched on the name of such occurances and it turns out to fall under systematic bias. I have not gone into what systematic bias yet.

It shows that forcing the model through the origin changes not just the numbers, but the direction in which it's wrong.

Something else I noticed is that, in the origin model, the "zero crossing" of the residuals happens somewhere in the middle of the data, which basically means the line pivots at the origin and swings away from the real trend for most of the range.

This pivoting also changes leverage patterns, extreme x points now have a lot of control over the slope, which makes the mid-range points consistently mispredicted. This is similar to a kind of leverage.

In the intercept model, leverage seems more evenly spread, so no single region of x can hijack the fit. But in the origin model, the higher valued x values if moved as a whole can drastically change the line, whereas a set of points moved in the intercept case won't drastically change the line as all the other sets still align in the same way and the averages prevent huge changes.

I also spotted a variance issue, in the origin model, residuals get larger in magnitude as x increases (heteroscedasticity). Heteroscedasticity just means that the errors don't stay at a constant level, it changes according to the x value.

This wasn't obvious from SSE alone, but in the plot, the 'spreading out' at higher x is clear. That means the model isn't equally reliable everywhere — predictions for large x are especially unstable and yield values with large errors.

Finally, I realised that the intercept itself isn't just a number at x=0, it's anchoring the whole regression so the slope is calculated in a way that reflects the actual center of the data cloud.

Removing it doesn't just shift the line vertically; it actually rotates the line and changes the balance between fitting low x and high x values.

This explains why the origin model fails in ways that aren't obvious from just looking at the slope or other mathematical quantities.

In short, while the intercept model is numerically better, it also preserves a random, homoscedastic error structure, avoids systematic bias across x, keeps residuals tightly centered, and maintains a slope supported by most of the data.

The origin-constrained model, despite looking like a valid straight line on paper, ends up being consistently and systematically wrong in ways that only become clear when you inspect residual plots and leverage patterns.

5 What I Learnt?

Mainly its the technical aspect that i learnt from this assignment. I learnt how to do simple data analysis and use formulae in excel to calculate other data from sample data.

I learnt how to create scatter plots and histograms and how to enter data into it. I learnt a little bit of LaTeX too.

Course Work related, I learnt how we can separate two different regression lines just using their SSE, Skewness, Kurtosis, R^2, etc. Rather than relying on my visualization, I have learnt how to decode how a pattern might look based on these mathematical notions.

I learnt how important the intercept for any regression curve it. If the intercept changes, its not as though the slope remains constant and the line shifts, the line changes its direction itself. So, in a way, if the direction is an important thing needed for prediction , then getting the intercept correct is highly important.

I also learnt about leverage where in wrong models, some sections of data if changed, can cause huge changes in the regression line. This mainly comes up due to the idea of heteroscedasticity, where the error associated with a particular value of x_i is not constant. This means predictions made will be off by huge margins at some place while little for some, ie: the pattern is not similar for all points.

One of the most important points I learnt was what each and every mathematical term meant, for example what is MAE, why is it needed, Kurtosis, how does it change, if it changes, what are the implications on the error histogram, etc.

Another thing I learnt the hard was this. One time, I paused my work in between and saved my file as a .csv file. But .csv files only store raw data, hence the formula i used to populate the cells is not visible for those cells I had already filled before saving. Only the numbers are visible, when I click the cells. Also, .csv files also don't save the plots generated. So for some columns of the .xlsx file I made, the formulae are not even visible.