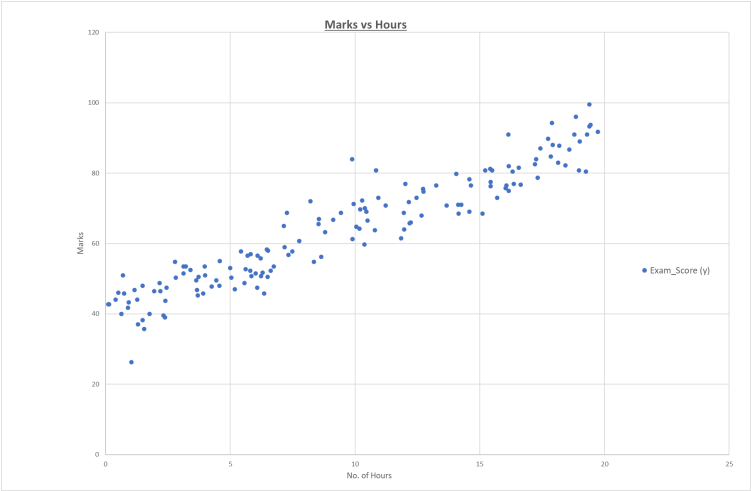
Part B

Question 1

The spreadsheet contains 150 samples of data containing the number of hours slept by the student and the marks scored in a test out of 100. Both the data fields (i.e. marks and hours) are ratio type data because they do contain a sensible true zero value and are continuous in nature

Question 2



The scatter plot as you can see above shows a clear positive linear trend between marks vs hours. The relation seems to be strong since the points are not that scattered and are close to each other and the spread of the points also seems to be the same throughout. By this I mean that the vertical spread / difference between points at a particular x is almost the same at an earlier vs a later x. This means that the variance of the data points is almost constant throughout since varaince is essentially the measure of how spread out the data is from the mean (which would be the line of best fit in this case). All of these observations suggest that Simple Regression Model is the appropriate choice for modeeling these two variables.

Question 3

β_1	=(M4-(K4*L4))/(N4-K4^2)	β_0	=(N4*L4 -M4*K4)/(N4-K4^2)

After applyong these formulae as shown the final value of coefficiemts is:

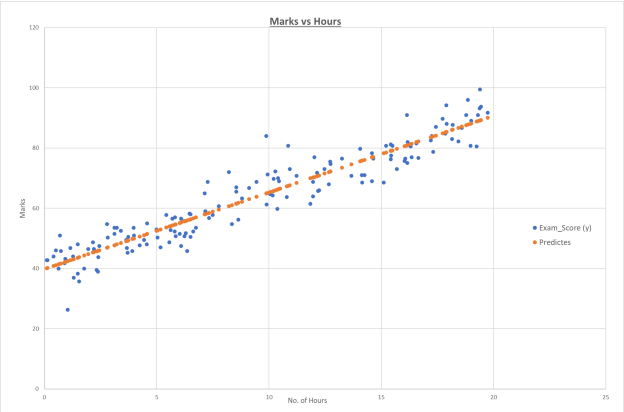
$$\beta_1 = 2.544, \beta_0 = 39.815$$

So substituing these values into the equation we get:

$$y_i = 39.815 + 2.544 \cdot x_i$$

This particular equation was then used to generate predicted values for each datapoint x and were stored in the next column in the spreadsheet. Note that a positive β_1 means that the positive relationship as observed from the scatter plot indeed exists and the score increases with number of hours.

Question 5

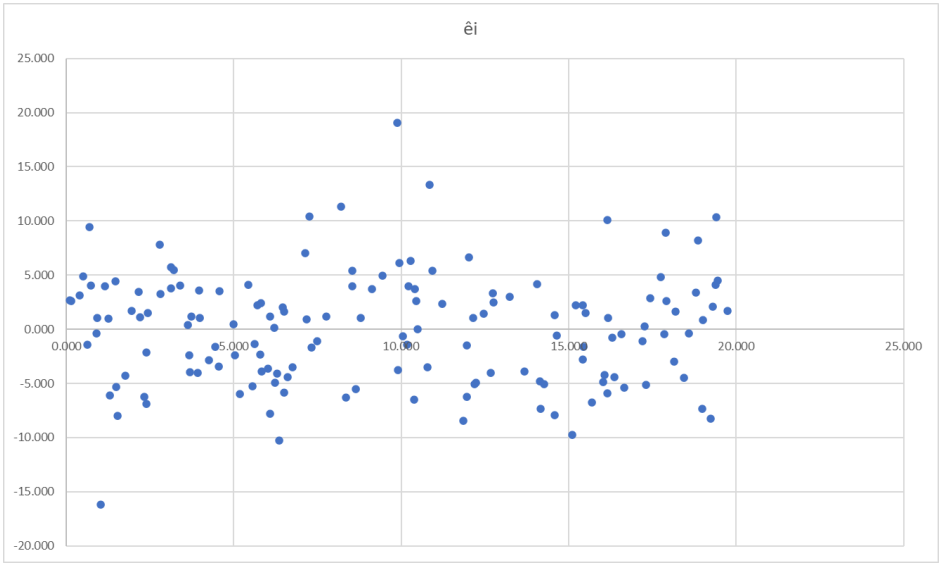


The plot above shows the original data points in blue with the predicted data points (in orange) by the model. The predicted values lie on a straight line (line of best fit) and the spread of the data can now be seen clearly to be not a lot and be tight. There are some points in anomaly that lie far away, but they are very few.

Question 6

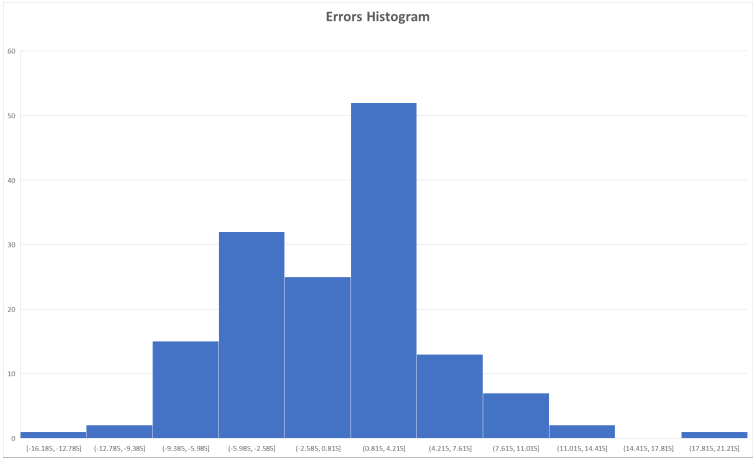
- SSE is the total squared difference between the actual and the predicted values. Its difficult to comment on the SSE because a larger number of data points will have a larger SSE so it can only be used to compare between two data sets of the same number of samples.
- Now MSE is SSE / N and hence it makesit a parameter that can be used for analysis. The downside of using MSE is that it is sensitive to large errors, as it squares the differences. But also if the residuals are less than 1, squaring them makes then smaller, not bigger, so MSE can be misleading, and if they are very small you might hit machine precision and be unable to calculate the MSE. For this data the MSE is 26.40 which means that the average of the square of errors is that particular value.
- RMSE is the square root of MSE. So it converts back the metric to its original units. It can sort of be interpreted as the standard difference between predicietd and actual errors. So on average the model's predictions are off by the RMSE value that is 5.14 marks. Because all are dependant on SSE, all three parameters are sensitive to outliers.
- MAE is not as sensitive to outliers and is just the mean of the errors. So to choose between which metric to use, RMSE is usually used when we wish to penalize outliers in data and large errors are undesirable. On the other hand if data has many outliers and all errors should be treated equally, MAE is preferred over RMSE.

Question 7



The plot above plots the residuals against the x values. Ideally the residuals should be random and not follow any pattern as such. If they did have any additional patterns it just is an indicator that there is an additional structure / feature that the model has not been able to capture and that the model is not appropriate. The plot above has random data points which confirms that the model is appropriate.

Question 8



The graph seems to be not centered at 0 because the mode is in the 0.8 to 4.2 bin which is not ideal since it means that the distribution is positively skewed. Also the tail seems to be very long that is the distribution should have been a little more spread out from the center. It also means that the model produces more positive errors. This means that the actual marks are greater than teh predicted marks and that the model is **underpredicting** than it should.

Question 9

Kurtosis is a statistical measure that describes the tailedness of a probability distribution which means how close the data points are to the center of the curve and how well are they spread out. Kurtosis value

for normal distribution is ideally 3 but the built in functions in excel calculate the relative value considering it to be 0 for normal distribution. Hence ideally a value closer to 0 indicated similarity to normal curves. **Skewness** refers to the symmetry/ asymmetry in the curve and whether the values are evenly distributed or not. So ideally it must be zero but the graph we have got seems to have a positive skewness.

Question 10

The R^2 value measures the proprtion of the variance in the Exam marks that is explained by the number of hours. The calculated R^2 value is 0.895. This means that the 89.5% of the varaibilty in scores can be accounted by the number of hours they studied. The remaning percentage is affected by other factors that have not been accounted for. The value of R^2 insidicates a very string goodness of fit and confirms a strong relationship between the two factors. R^2 cant detect if a model is biased or not towards a particular kinds of error and just sees if the points are very close to the line on average. Thus the R^2 with the residual of errors is important to comment on the reliabilty of the model. But since both R^2 and the residual plots are in favor of the model, we can safely conclude that the model is a perofming well.

Question 11

We force the intercept to be zero when there is some physical relevance to the variables and we know that if x=0 then y will also be 0. Some examples would include Ohm's Law where its a fact that if I=0 then V=0 too. On comparison with the earlier model its clear that forcing the intercept to be zero can not be done here. The residual seems to be linearly decreasing and the sum of square of errors is huge.