DS203: Assignment 1

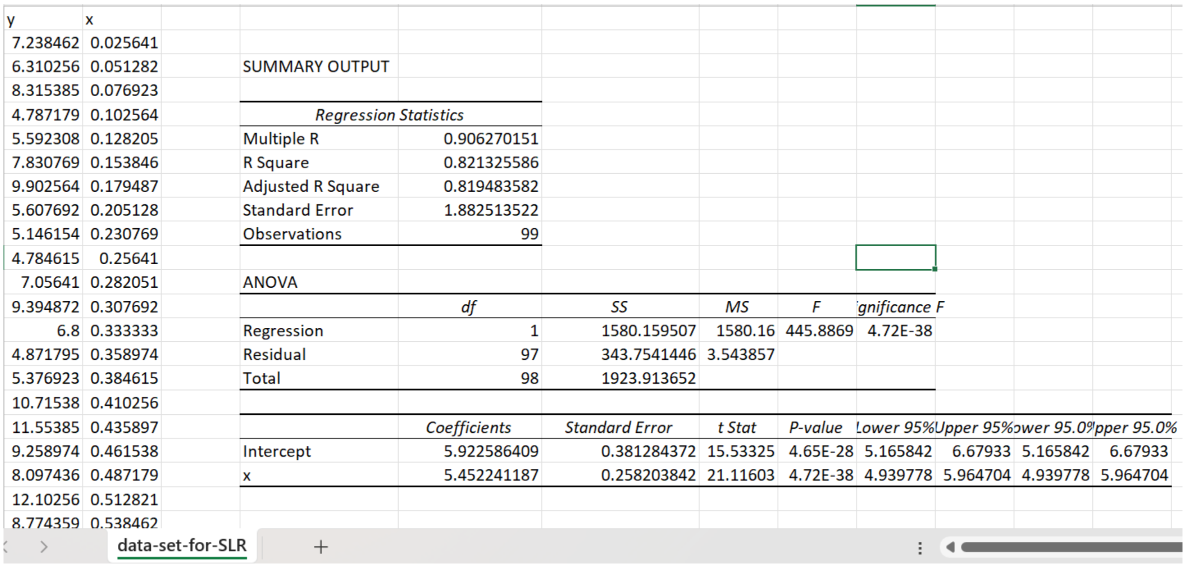
Sahil Rakesh Wani
210070070

January 21, 2024

The assignment contains two parts: Part-A and Part-B

1 Part - A

This is the final SLR model summary created using the inbuilt **Linear Regression functionality** in the **Data Analysis Toolpack** in MS Excel.



y	x
7.238462	0.025641
6.310256	0.051282
8.315385	0.076923
4.787179	0.102564
5.592308	0.128205
7.830769	0.153846
9.902564	0.179487
5.607692	0.205128
5.146154	0.230769
4.784615	0.25641
7.05641	0.282051
9.394872	0.307692
6.8	0.333333
4.871795	0.358974
5.376923	0.384615
10.71538	0.410256
11.55385	0.435897
9.258974	0.461538
8.097436	0.487179
12.10256	0.512821
8.774359	0.538462

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.906270151
R Square	0.821325586
Adjusted R Square	0.819483582
Standard Error	1.882513522
Observations	99

ANOVA

	df	*SS*	*MS*	*F*	*gnificance F*
Regression	1	1580.159507	1580.16	445.8869	4.72E-38
Residual	97	343.7541446	3.543857		
Total	98	1923.913652			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
Intercept	5.922586409	0.381284372	15.53325	4.65E-28	5.165842	6.67933	5.165842	6.67933
x	5.452241187	0.258203842	21.11603	4.72E-38	4.939778	5.964704	4.939778	5.964704

Figure 1: Regression Model Summary using inbuilt functionality

2 Part - B

2.1 Question 7

Below is the **y v/s x** scatter plot for the train data.

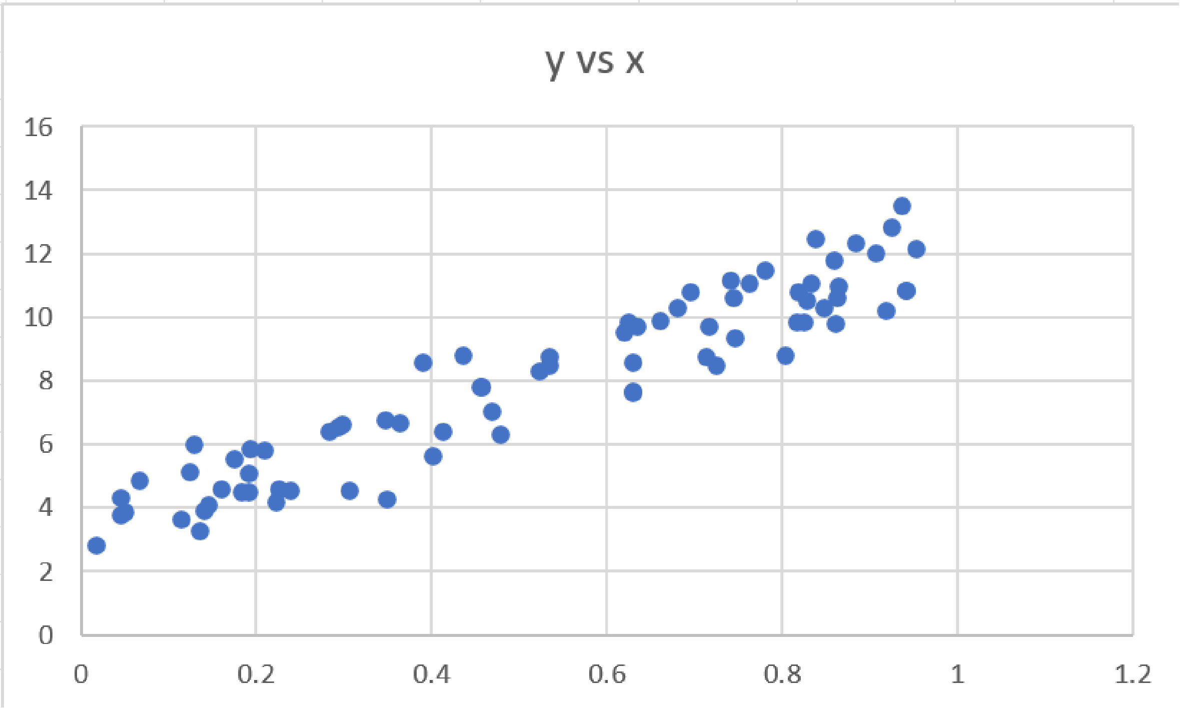
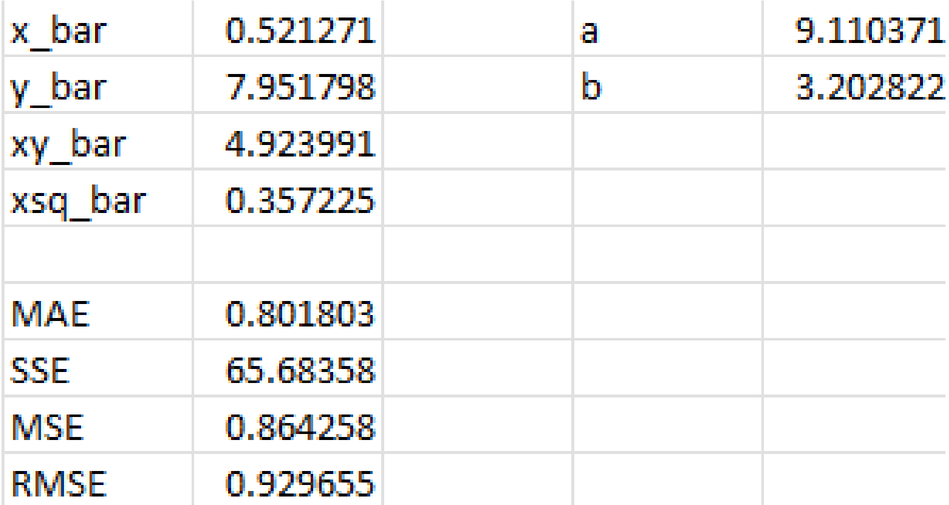


Figure 2: **y v/s x** scatter plot for training data

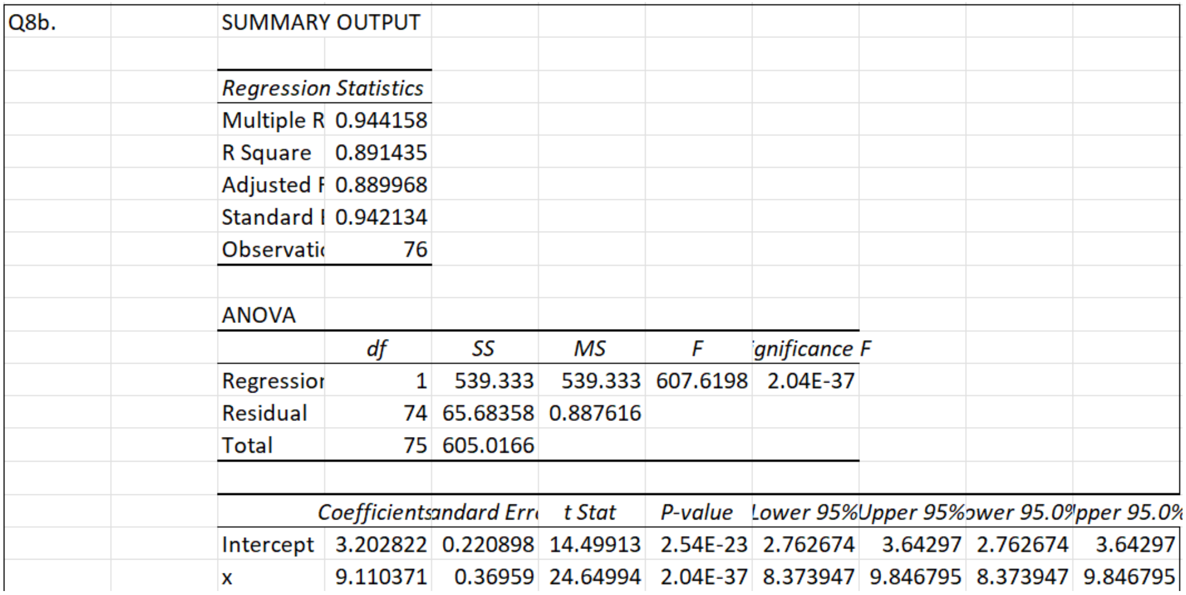
2.2 Question 8

For the training dataset:



x_bar	0.521271		a	9.110371
y_bar	7.951798		b	3.202822
xy_bar	4.923991			
xsq_bar	0.357225			
MAE	0.801803			
SSE	65.68358			
MSE	0.864258			
RMSE	0.929655			

Figure 3: a and b using closed form equations for train_data



Q8b.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.944158
R Square	0.891435
Adjusted F	0.889968
Standard I	0.942134
Observatio	76

ANOVA

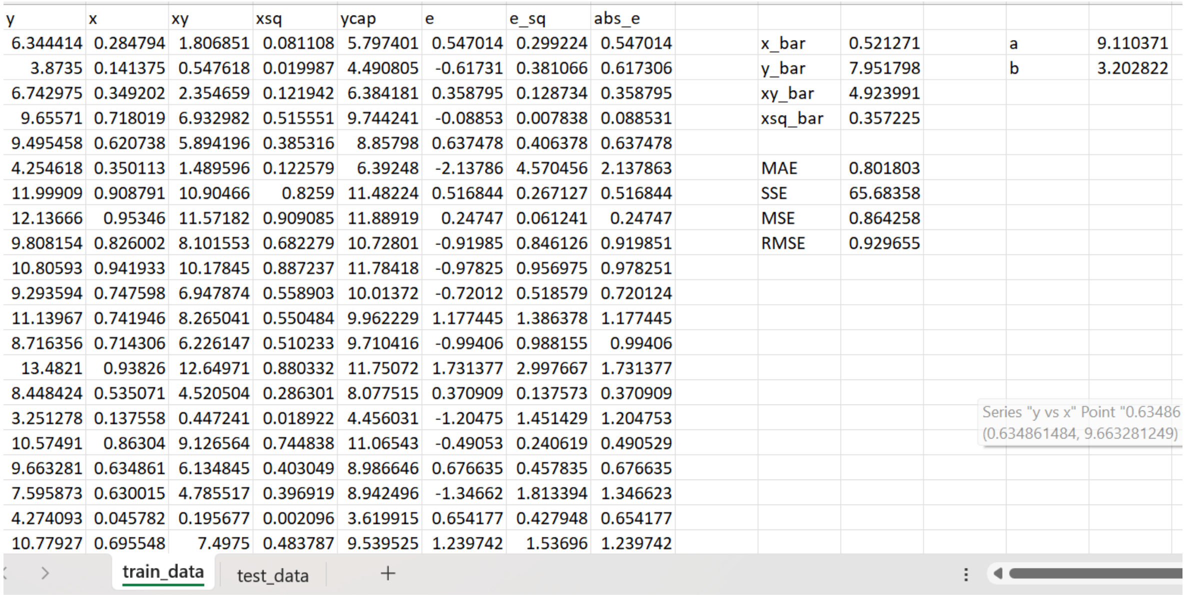
	df	*SS*	*MS*	*F*	*ignificance F*
Regressior	1	539.333	539.333	607.6198	2.04E-37
Residual	74	65.68358	0.887616		
Total	75	605.0166			

	Coefficients	*andard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
Intercept	3.202822	0.220898	14.49913	2.54E-23	2.762674	3.64297	2.762674	3.64297
x	9.110371	0.36959	24.64994	2.04E-37	8.373947	9.846795	8.373947	9.846795

Figure 4: Regression summary using in-built functionalities for train_data

2.3 Question 9

Below are calculated values using the training dataset:



y	x	xy	xsq	ycap	e	e_sq	abs_e
6.344414	0.284794	1.806851	0.081108	5.797401	0.547014	0.299224	0.547014
3.8735	0.141375	0.547618	0.019987	4.490805	-0.61731	0.381066	0.617306
6.742975	0.349202	2.354659	0.121942	6.384181	0.358795	0.128734	0.358795
9.65571	0.718019	6.932982	0.515551	9.744241	-0.08853	0.007838	0.088531
9.495458	0.620738	5.894196	0.385316	8.85798	0.637478	0.406378	0.637478
4.254618	0.350113	1.489596	0.122579	6.39248	-2.13786	4.570456	2.137863
11.99909	0.908791	10.90466	0.8259	11.48224	0.516844	0.267127	0.516844
12.13666	0.95346	11.57182	0.909085	11.88919	0.24747	0.061241	0.24747
9.808154	0.826002	8.101553	0.682279	10.72801	-0.91985	0.846126	0.919851
10.80593	0.941933	10.17845	0.887237	11.78418	-0.97825	0.956975	0.978251
9.293594	0.747598	6.947874	0.558903	10.01372	-0.72012	0.518579	0.720124
11.13967	0.741946	8.265041	0.550484	9.962229	1.177445	1.386378	1.177445
8.716356	0.714306	6.226147	0.510233	9.710416	-0.99406	0.988155	0.99406
13.4821	0.93826	12.64971	0.880332	11.75072	1.731377	2.997667	1.731377
8.448424	0.535071	4.520504	0.286301	8.077515	0.370909	0.137573	0.370909
3.251278	0.137558	0.447241	0.018922	4.456031	-1.20475	1.451429	1.204753
10.57491	0.86304	9.126564	0.744838	11.06543	-0.49053	0.240619	0.490529
9.663281	0.634861	6.134845	0.403049	8.986646	0.676635	0.457835	0.676635
7.595873	0.630015	4.785517	0.396919	8.942496	-1.34662	1.813394	1.346623
4.274093	0.045782	0.195677	0.002096	3.619915	0.654177	0.427948	0.654177
10.77927	0.695548	7.4975	0.483787	9.539525	1.239742	1.53696	1.239742

x_bar	0.521271	a	9.110371
y_bar	7.951798	b	3.202822
xy_bar	4.923991		
xsq_bar	0.357225		
MAE	0.801803		
SSE	65.68358		
MSE	0.864258		
RMSE	0.929655		

Figure 5: Calculations in Excel using train data

Quantities	Values
MAE	0.801803
SSE	65.68358
MSE	0.864258
RMSE	0.929655

Table 1: Statistics of the training data sample

The MAE, MSE and RMSE are quite close to the value 1, and the y data values range from approx. 2 to 14. This indicates that the error has been quite small and linear regression is a decent model for the given dataset.

2.4 Question 10

Blue dots represent $\boldsymbol{y}$ ***vs*** $\boldsymbol{x}$, and **Orange** dots represent $\hat{y}$ ***vs*** $\boldsymbol{x}$

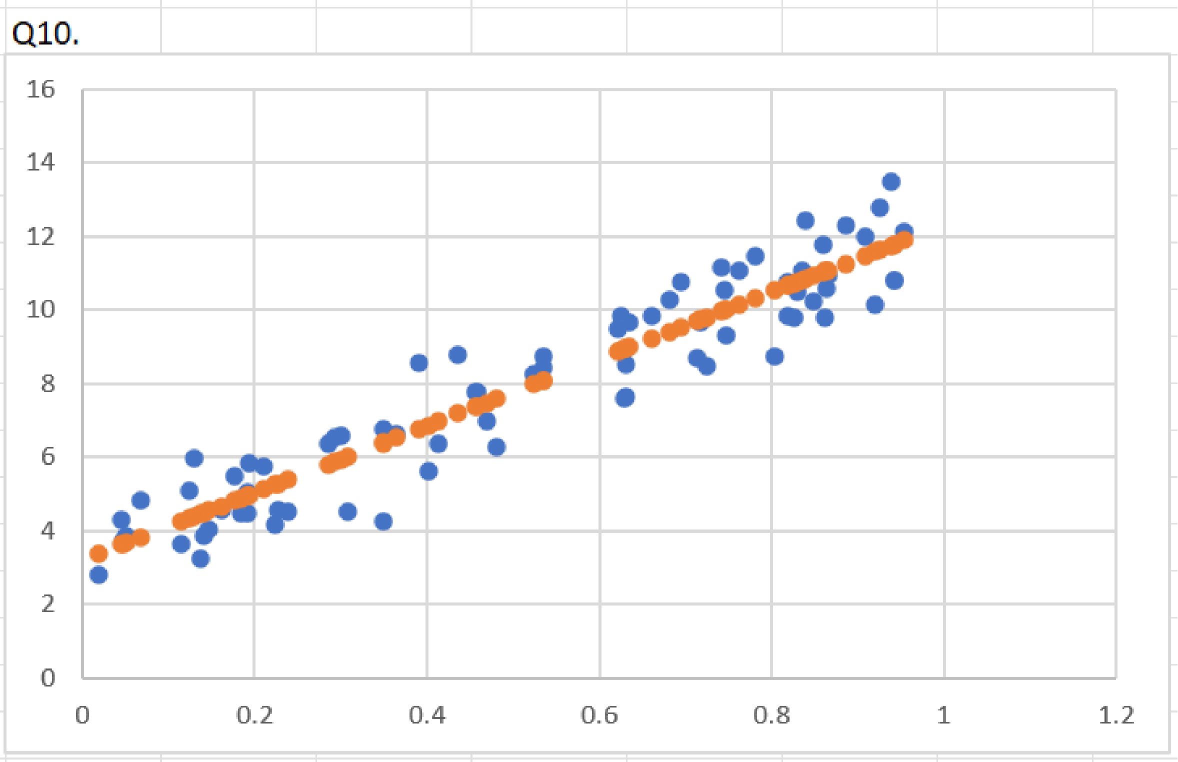


Figure 6: Superimposed scatter plot

The predicted ycap values represent the linear regression line. As our sample data is pretty close to be represented by a line, linear regression seems to be a good choice as a model.

3 Question 11

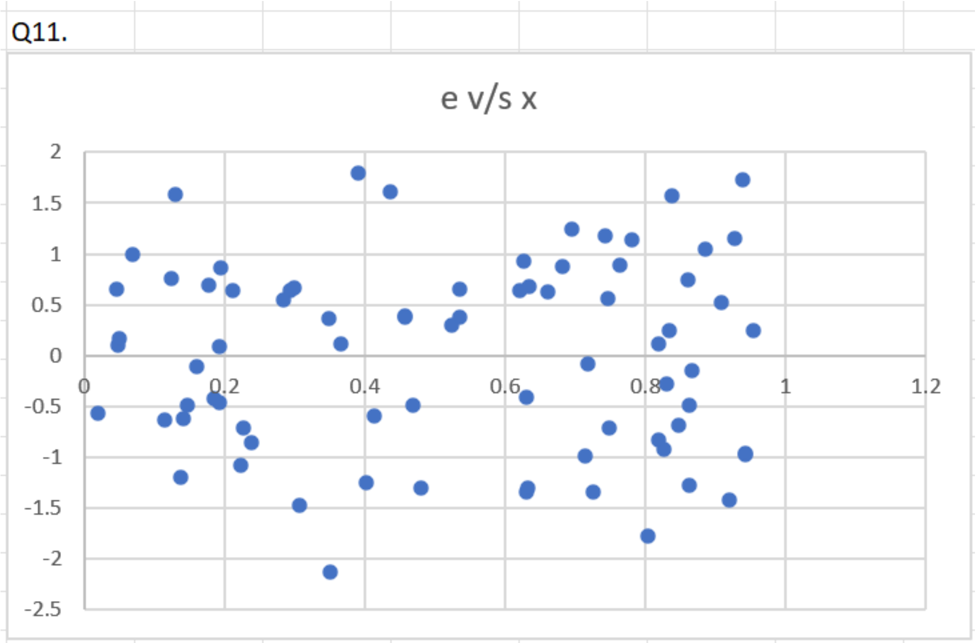
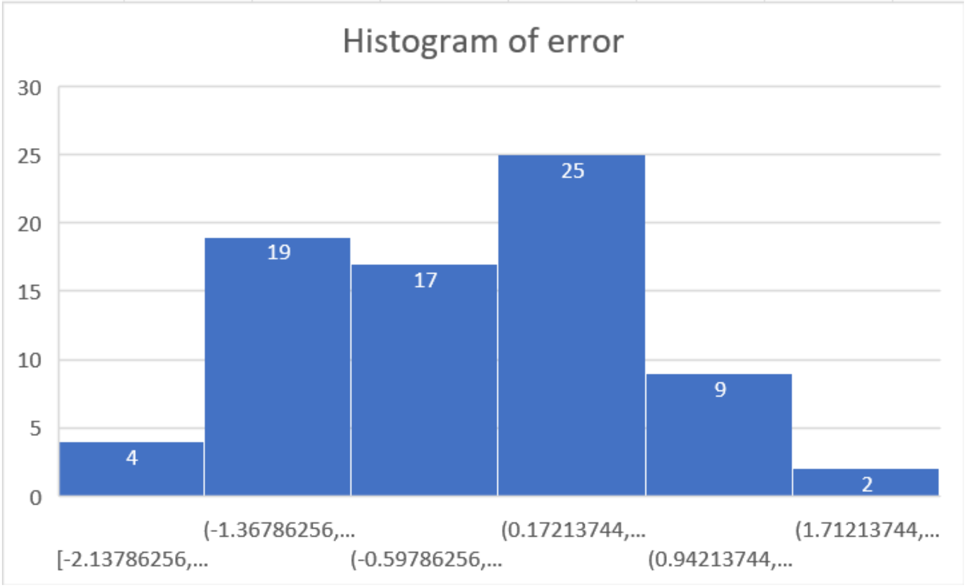


Figure 7: Scatter plot of error vs x

As the scatter plot of the error seems random and no particular trend is visible, it can be concluded that our model is fitting well to the sample data.

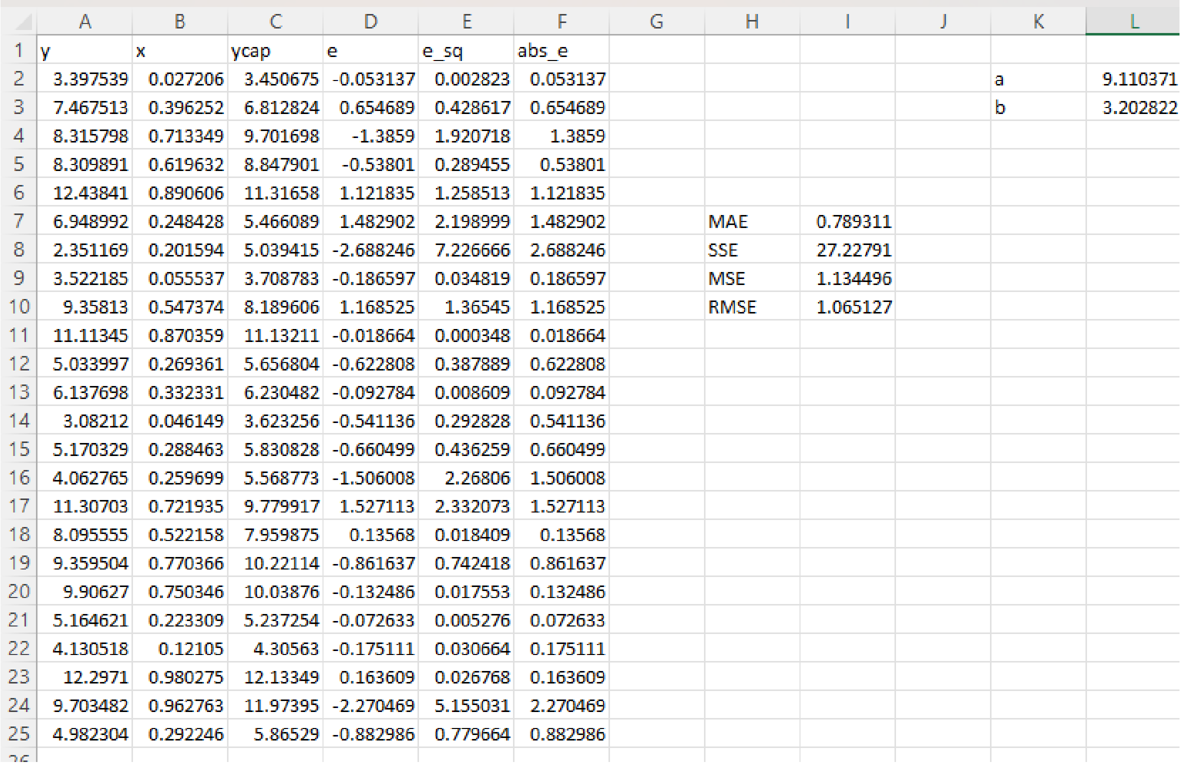
4 Question 12



As the distribution seeems fairly close to a normal distribution, our model is fitting decently to the sample data.

5 Question 13 : Calculating statistics for the test data

5.1 General Statistics



	A	B	C	D	E	F	G	H	I	J	K	L
1	y	x	ycap	e	e_sq	abs_e						
2	3.397539	0.027206	3.450675	-0.053137	0.002823	0.053137					a	9.110371
3	7.467513	0.396252	6.812824	0.654689	0.428617	0.654689					b	3.202822
4	8.315798	0.713349	9.701698	-1.3859	1.920718	1.3859						
5	8.309891	0.619632	8.847901	-0.53801	0.289455	0.53801						
6	12.43841	0.890606	11.31658	1.121835	1.258513	1.121835						
7	6.948992	0.248428	5.466089	1.482902	2.198999	1.482902		MAE	0.789311			
8	2.351169	0.201594	5.039415	-2.688246	7.226666	2.688246		SSE	27.22791			
9	3.522185	0.055537	3.708783	-0.186597	0.034819	0.186597		MSE	1.134496			
10	9.35813	0.547374	8.189606	1.168525	1.36545	1.168525		RMSE	1.065127			
11	11.11345	0.870359	11.13211	-0.018664	0.000348	0.018664						
12	5.033997	0.269361	5.656804	-0.622808	0.387889	0.622808						
13	6.137698	0.332331	6.230482	-0.092784	0.008609	0.092784						
14	3.08212	0.046149	3.623256	-0.541136	0.292828	0.541136						
15	5.170329	0.288463	5.830828	-0.660499	0.436259	0.660499						
16	4.062765	0.259699	5.568773	-1.506008	2.26806	1.506008						
17	11.30703	0.721935	9.779917	1.527113	2.332073	1.527113						
18	8.095555	0.522158	7.959875	0.13568	0.018409	0.13568						
19	9.359504	0.770366	10.22114	-0.861637	0.742418	0.861637						
20	9.90627	0.750346	10.03876	-0.132486	0.017553	0.132486						
21	5.164621	0.223309	5.237254	-0.072633	0.005276	0.072633						
22	4.130518	0.12105	4.30563	-0.175111	0.030664	0.175111						
23	12.2971	0.980275	12.13349	0.163609	0.026768	0.163609						
24	9.703482	0.962763	11.97395	-2.270469	5.155031	2.270469						
25	4.982304	0.292246	5.86529	-0.882986	0.779664	0.882986						

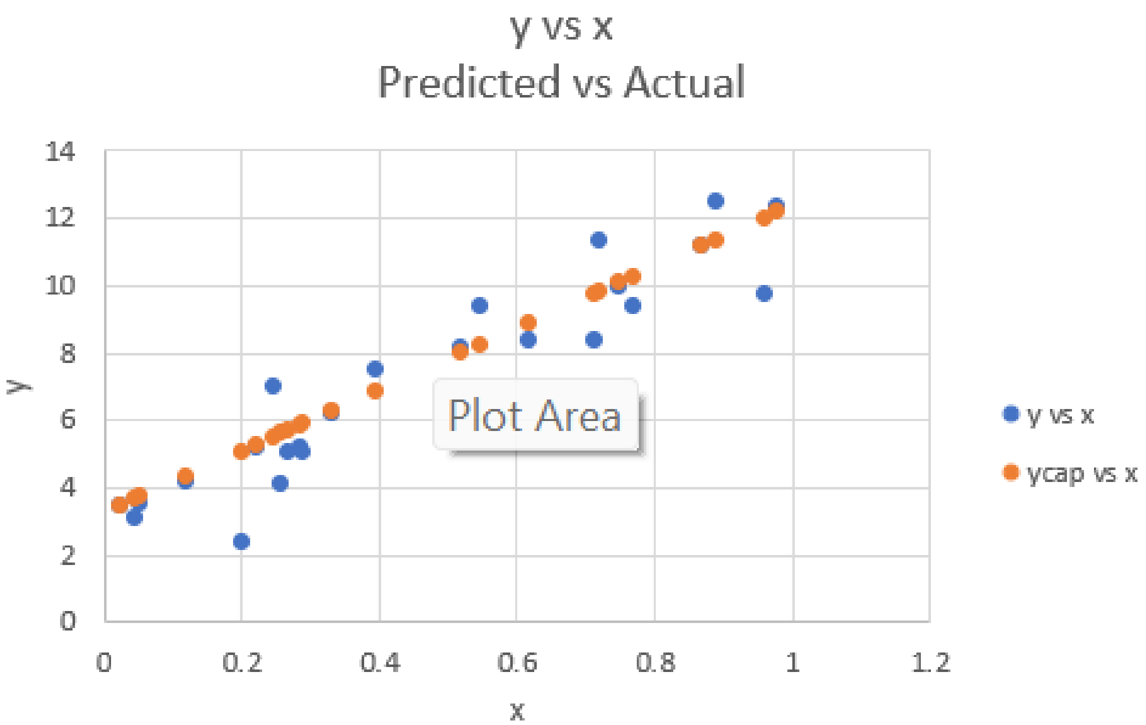
Figure 8: Calculations in Excel

Quantities	Values
MAE	0.78
SSE	27.22
MSE	1.13
RMSE	1.06

Table 2: Statistics of the test data sample

The mean errors are still quite close to 1, and the y data values range approx. from 2 to 13. This indicates that the error has been quite small and our model is working as expected.

5.2 Scatter Plots of the training dataset



The predicted ycap values represent the linear regression line. As our test data is pretty close to be represented by a line, our linear regression model seems to be working well.

5.3 Scatter Plot of error vs x

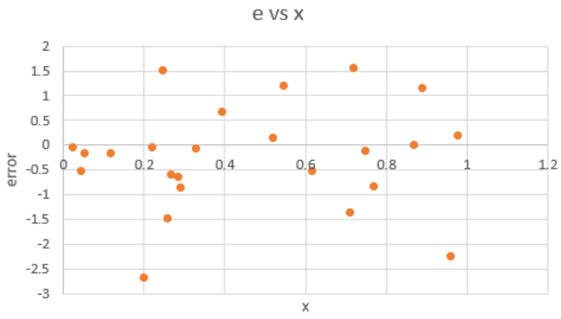


Figure 9: Scatter Plot for training data

It can be observed that there is no significant trend in the error scatter plot. We conclude that our model fits well on the training data as well.

5.4 Histogram of error for training data

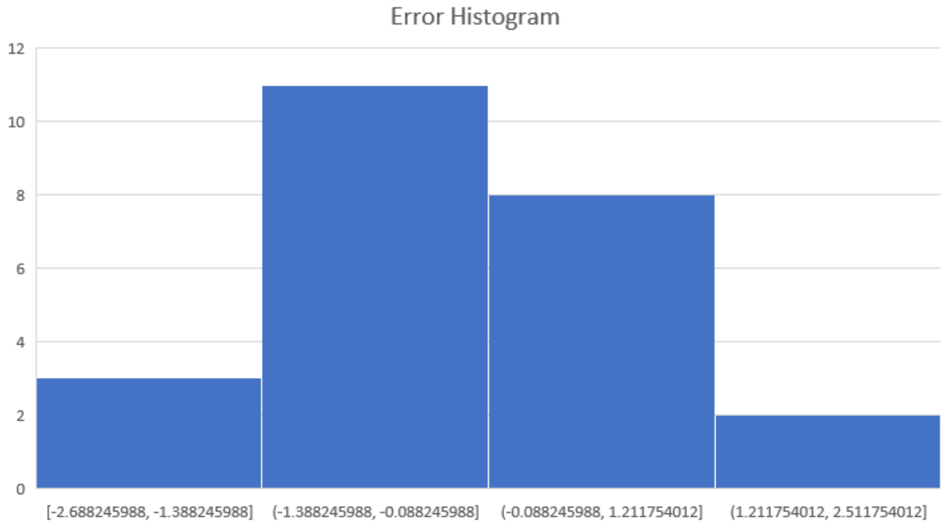


Figure 10: Histogram of e for training data

The plot seems close to the normal distribution, which is desired and indicates of a good model.

6 Comparison of error metrics and plots

The mean of the errors(MSE, RMSE) have increased slightly when we analyzed the test data, as compared to the train data. This indicates that the model performance has decreased slightly, though it is negligible.

The error plots of both the datasets have been random, with no significant trend visible. This is a good indicator of our model, and indicates that the model choice is appropriate for the given data. The error histograms show a normal distribution. This is also a good indicator of our model fitting the sample data.

The training error is also not very close to zero, thus ensuring that our model is not overfitting the training data.