# DS203: Assignment 1

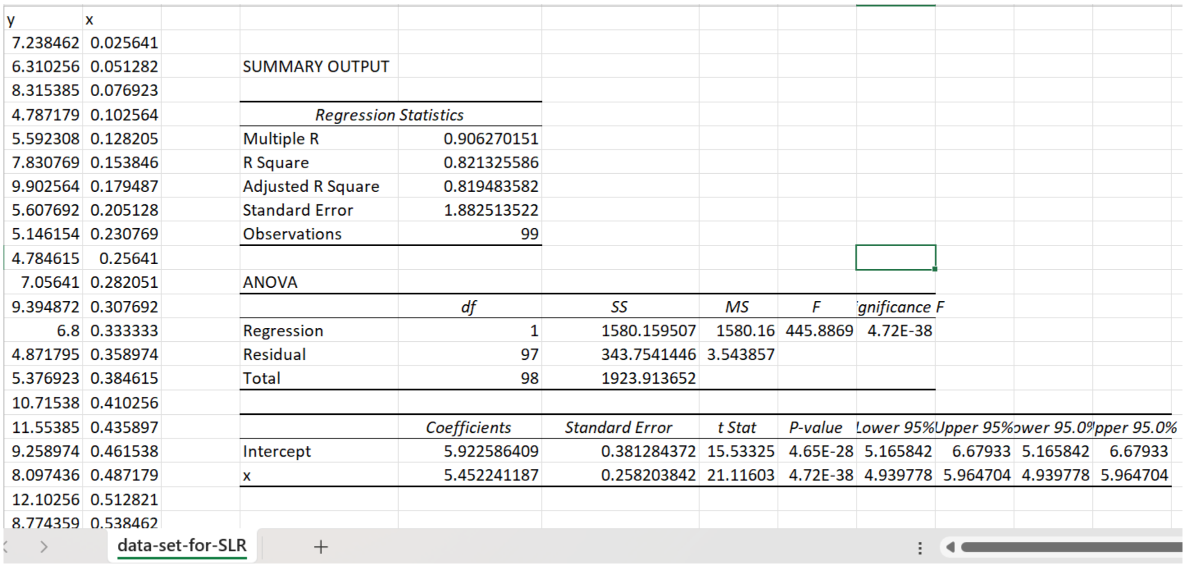
Sahil Rakesh Wani
210070070

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.

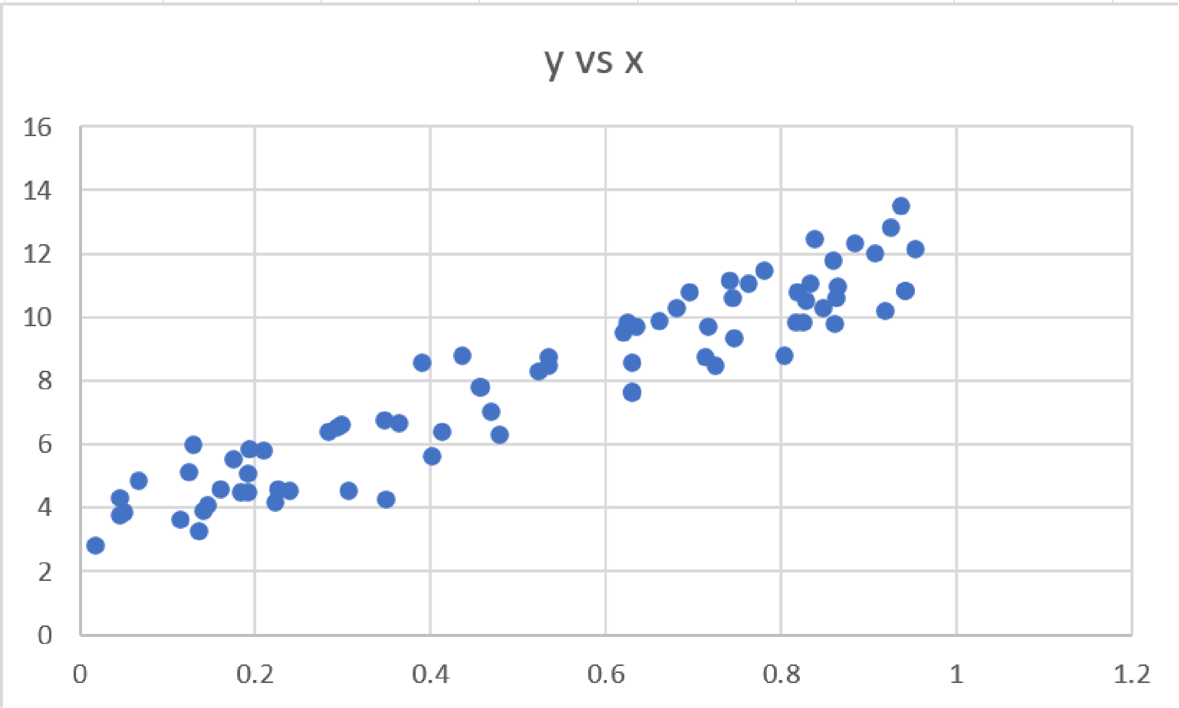
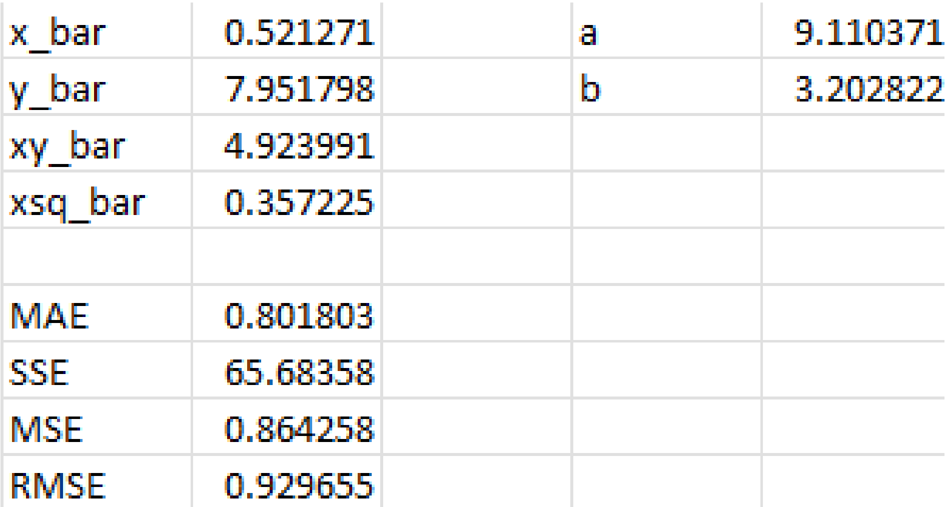


Figure 2: **y v/s x** scatter plot for training data

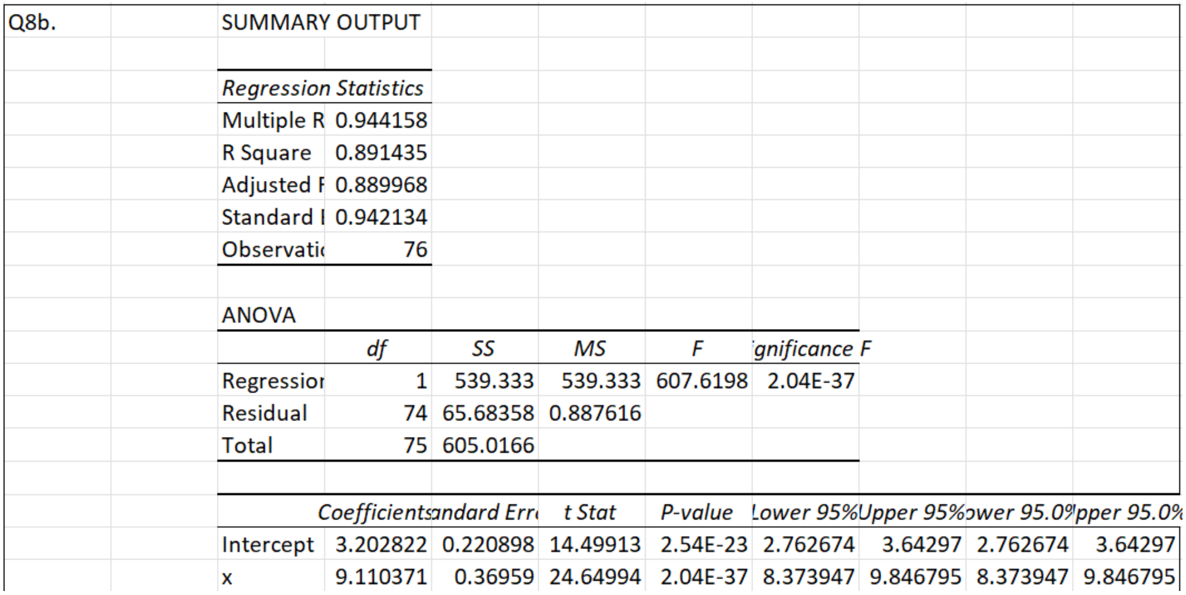
## 2.2 Question 8

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 |
| Observati | 76 |

ANOVA

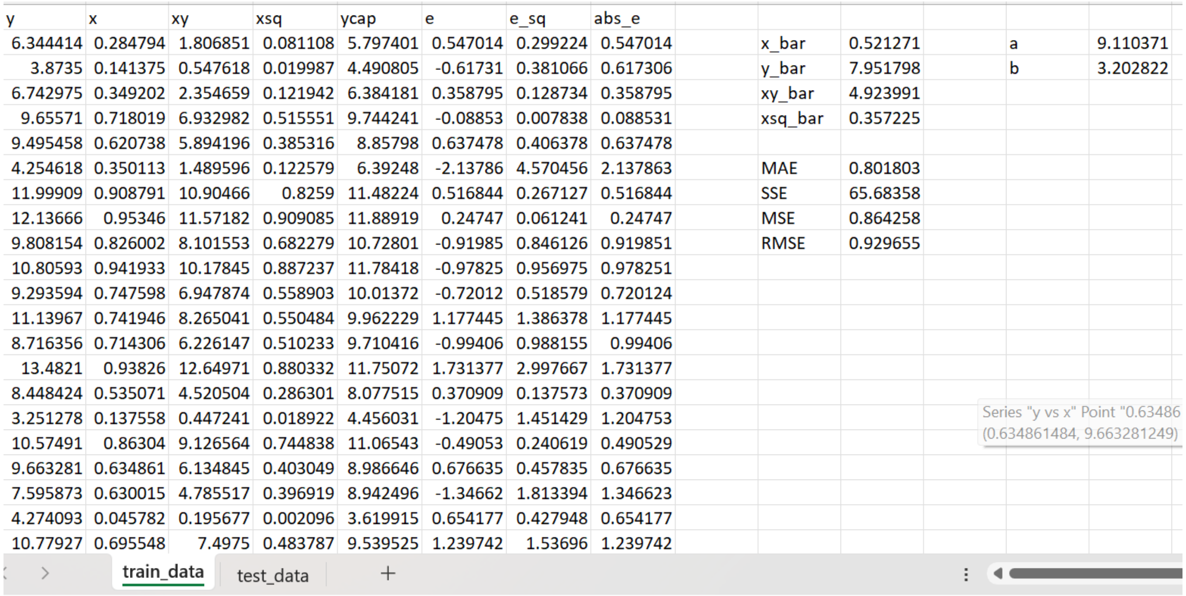
| | *df* | *SS* | *MS* | *F* | *gnificance F* |
|---|---|---|---|---|---|
| Regressior | 1 | 539.333 | 539.333 | 607.6198 | 2.04E-37 |
| Residual | 74 | 65.68358 | 0.887616 | | |
| Total | 75 | 605.0166 | | | |

| | *Coefficients* | *andard Erri* | *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 | | |

Series "y vs x" Point "0.63486
(0.634861484, 9.663281249)

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*** $\boldsymbol{x}$

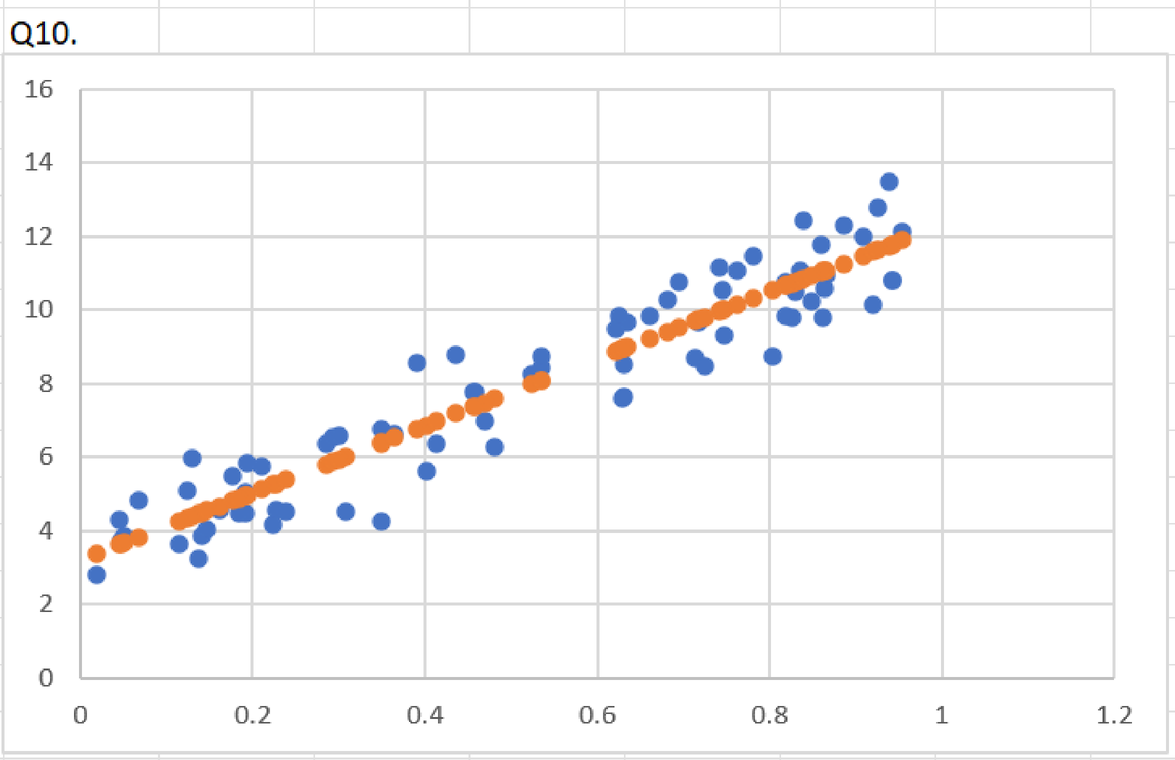


Figure 6: Superimposed scatter plot

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

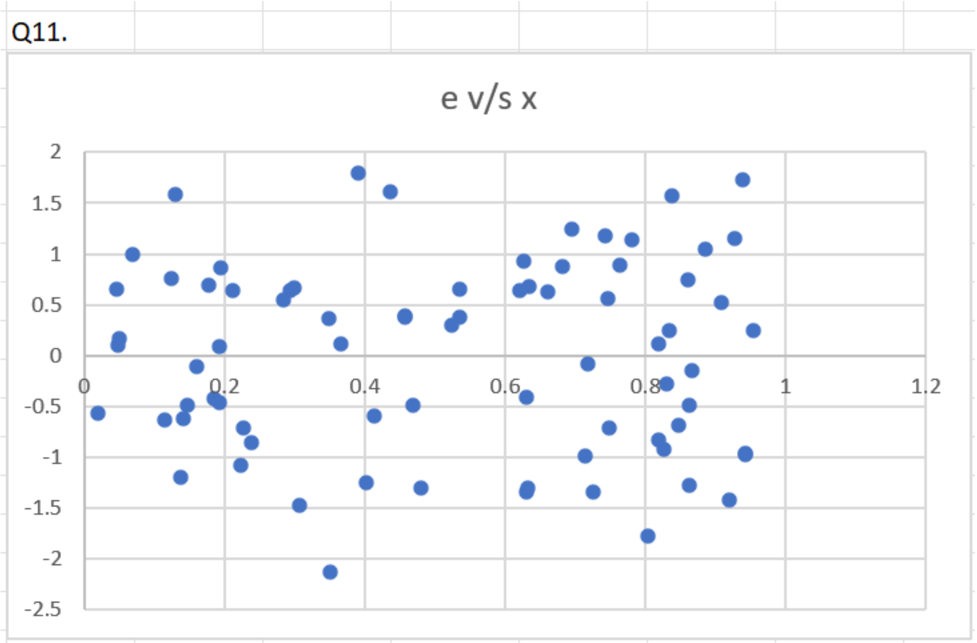
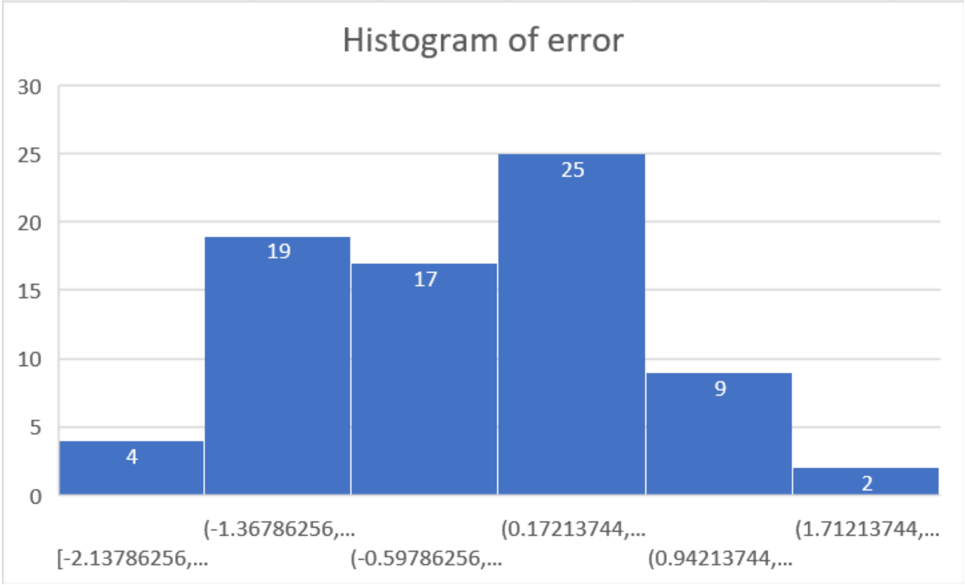


Figure 7: Scatter plot of error vs x

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.

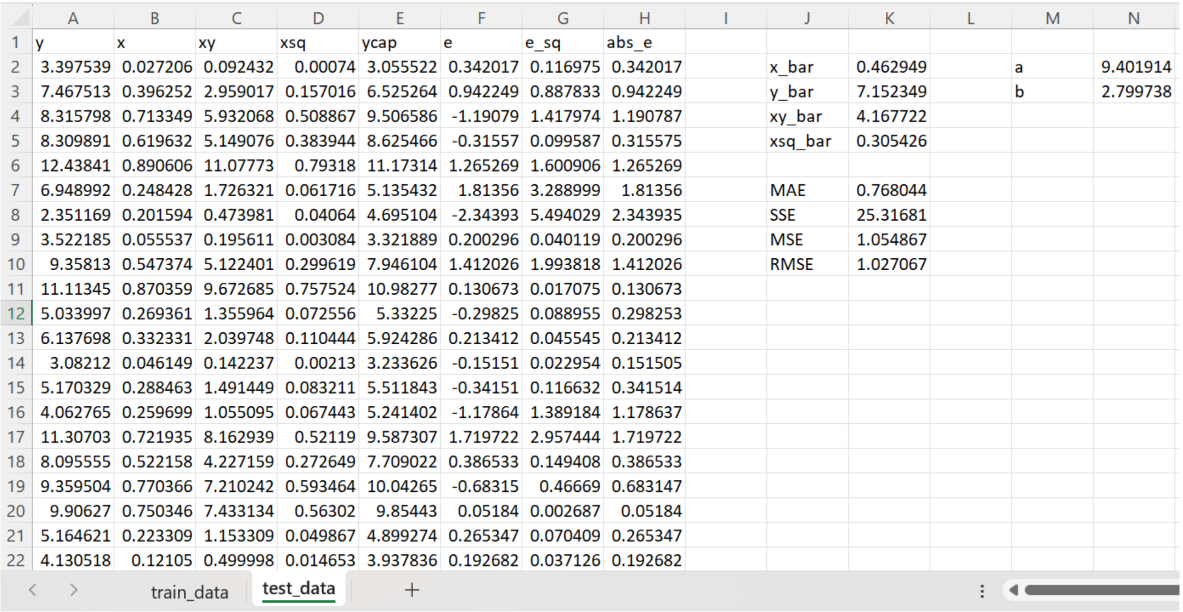
# 4 Question 12



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 | M | N |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | y | x | xy | xsq | ycap | e | e_sq | abs_e | | | | | | |
| 2 | 3.397539 | 0.027206 | 0.092432 | 0.00074 | 3.055522 | 0.342017 | 0.116975 | 0.342017 | | x_bar | 0.462949 | | a | 9.401914 |
| 3 | 7.467513 | 0.396252 | 2.959017 | 0.157016 | 6.525264 | 0.942249 | 0.887833 | 0.942249 | | y_bar | 7.152349 | | b | 2.799738 |
| 4 | 8.315798 | 0.713349 | 5.932068 | 0.508867 | 9.506586 | -1.19079 | 1.417974 | 1.190787 | | xy_bar | 4.167722 | | | |
| 5 | 8.309891 | 0.619632 | 5.149076 | 0.383944 | 8.625466 | -0.31557 | 0.099587 | 0.315575 | | xsq_bar | 0.305426 | | | |
| 6 | 12.43841 | 0.890606 | 11.07773 | 0.79318 | 11.17314 | 1.265269 | 1.600906 | 1.265269 | | | | | | |
| 7 | 6.948992 | 0.248428 | 1.726321 | 0.061716 | 5.135432 | 1.81356 | 3.288999 | 1.81356 | | MAE | 0.768044 | | | |
| 8 | 2.351169 | 0.201594 | 0.473981 | 0.04064 | 4.695104 | -2.34393 | 5.494029 | 2.343935 | | SSE | 25.31681 | | | |
| 9 | 3.522185 | 0.055537 | 0.195611 | 0.003084 | 3.321889 | 0.200296 | 0.040119 | 0.200296 | | MSE | 1.054867 | | | |
| 10 | 9.35813 | 0.547374 | 5.122401 | 0.299619 | 7.946104 | 1.412026 | 1.993818 | 1.412026 | | RMSE | 1.027067 | | | |
| 11 | 11.11345 | 0.870359 | 9.672685 | 0.757524 | 10.98277 | 0.130673 | 0.017075 | 0.130673 | | | | | | |
| 12 | 5.033997 | 0.269361 | 1.355964 | 0.072556 | 5.33225 | -0.29825 | 0.088955 | 0.298253 | | | | | | |
| 13 | 6.137698 | 0.332331 | 2.039748 | 0.110444 | 5.924286 | 0.213412 | 0.045545 | 0.213412 | | | | | | |
| 14 | 3.08212 | 0.046149 | 0.142237 | 0.00213 | 3.233626 | -0.15151 | 0.022954 | 0.151505 | | | | | | |
| 15 | 5.170329 | 0.288463 | 1.491449 | 0.083211 | 5.511843 | -0.34151 | 0.116632 | 0.341514 | | | | | | |
| 16 | 4.062765 | 0.259699 | 1.055095 | 0.067443 | 5.241402 | -1.17864 | 1.389184 | 1.178637 | | | | | | |
| 17 | 11.30703 | 0.721935 | 8.162939 | 0.52119 | 9.587307 | 1.719722 | 2.957444 | 1.719722 | | | | | | |
| 18 | 8.095555 | 0.522158 | 4.227159 | 0.272649 | 7.709022 | 0.386533 | 0.149408 | 0.386533 | | | | | | |
| 19 | 9.359504 | 0.770366 | 7.210242 | 0.593464 | 10.04265 | -0.68315 | 0.46669 | 0.683147 | | | | | | |
| 20 | 9.90627 | 0.750346 | 7.433134 | 0.56302 | 9.85443 | 0.05184 | 0.002687 | 0.05184 | | | | | | |
| 21 | 5.164621 | 0.223309 | 1.153309 | 0.049867 | 4.899274 | 0.265347 | 0.070409 | 0.265347 | | | | | | |
| 22 | 4.130518 | 0.12105 | 0.499998 | 0.014653 | 3.937836 | 0.192682 | 0.037126 | 0.192682 | | | | | | |

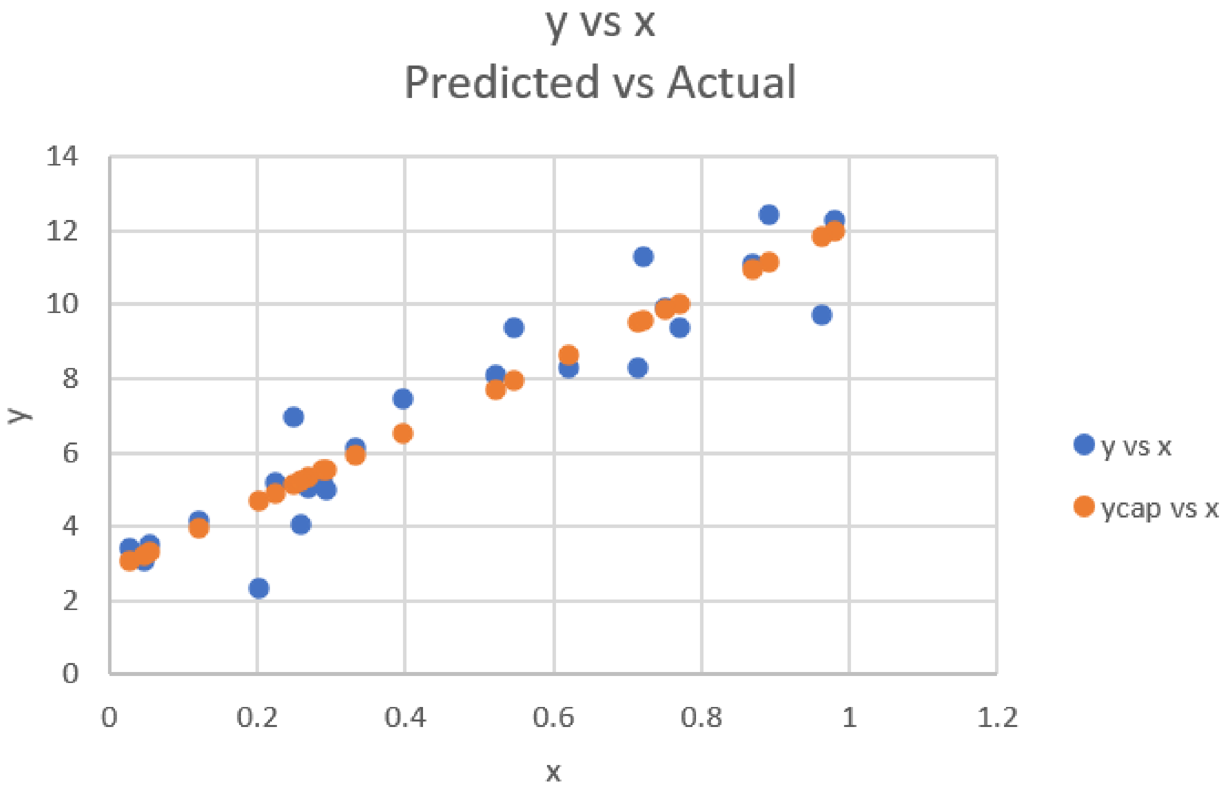
Figure 8: Calculations in Excel

| Quantities | Values |
|---|---|
| MAE | 0.76 |
| SSE | 25.31 |
| MSE | 1.05 |
| RMSE | 1.02 |

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

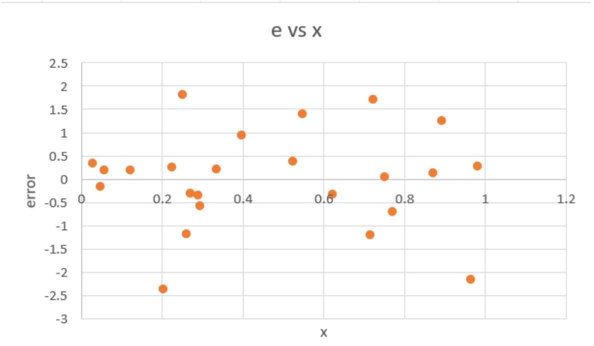


Figure 9: Scatter Plot for training data

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

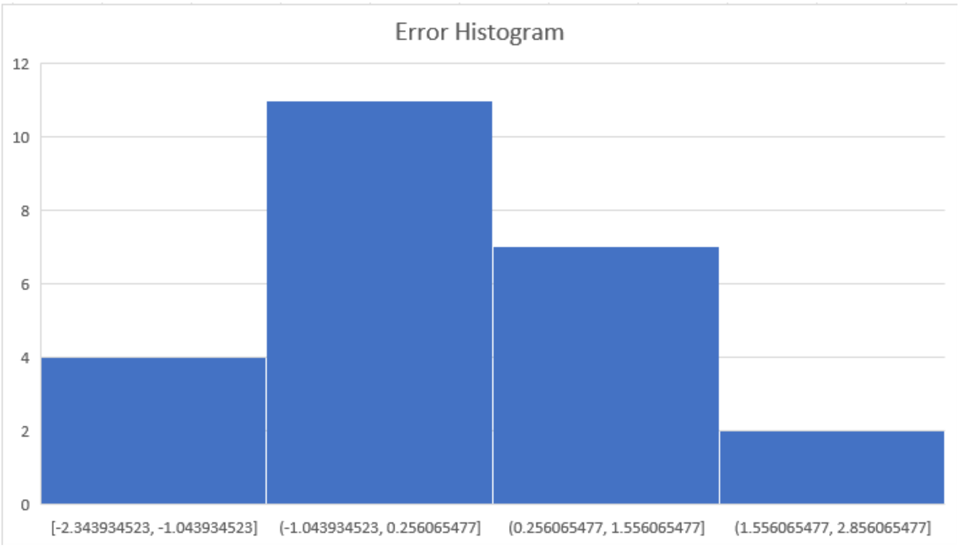


Figure 10: Histogram of e for training data

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.