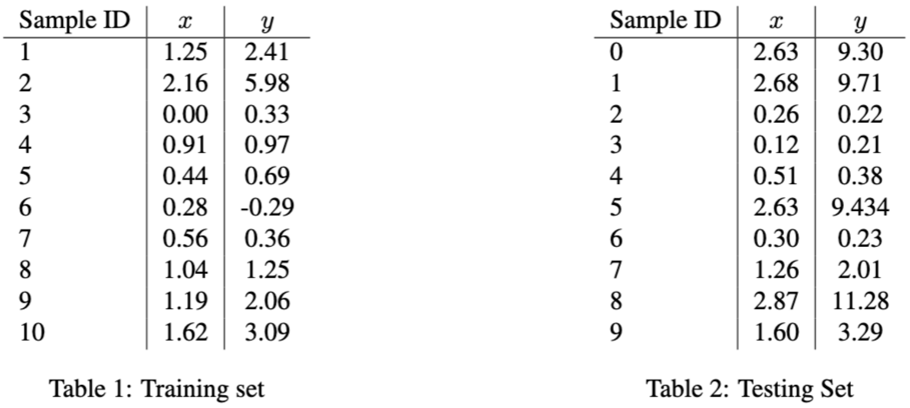
Mercury Mcindoe 85594505

1) Linear Regression



Sample ID	x	y
1	1.25	2.41
2	2.16	5.98
3	0.00	0.33
4	0.91	0.97
5	0.44	0.69
6	0.28	-0.29
7	0.56	0.36
8	1.04	1.25
9	1.19	2.06
10	1.62	3.09

Table 1: Training set

Sample ID	x	y
0	2.63	9.30
1	2.68	9.71
2	0.26	0.22
3	0.12	0.21
4	0.51	0.38
5	2.63	9.434
6	0.30	0.23
7	1.26	2.01
8	2.87	11.28
9	1.60	3.29

Table 2: Testing Set

Figure 1: Question 1 tables.

Please fit the linear model $\hat{Y} = f_\Theta(X)$ using **RSS** objective $J(\Theta) = \|\hat{Y} - Y\|_2^2$.

Given the RSS objective $J(\Theta) = \|\hat{Y} - Y\|_2^2$, we can obtain the analytical solution by the following procedure. $J(\Theta) = (X\Theta - Y)^T(X\Theta - Y) = \Theta^T X^T - Y^T X\Theta - \Theta^T X^T Y + Y^T Y$. $J''(\Theta) = 2X^T X > 0, J'(\Theta) = 2X^T(X\Theta - Y) = 0 \to \Theta^* = (X^T X)^{-1} X^T Y$.

(a) Calculate the (closed-form) analytic solution of the linear model $f_\Theta(x) = \theta_0 + \theta_1 x$. Then, use a scatter plot to plot the training data points and draw the fitted line on the same figure.

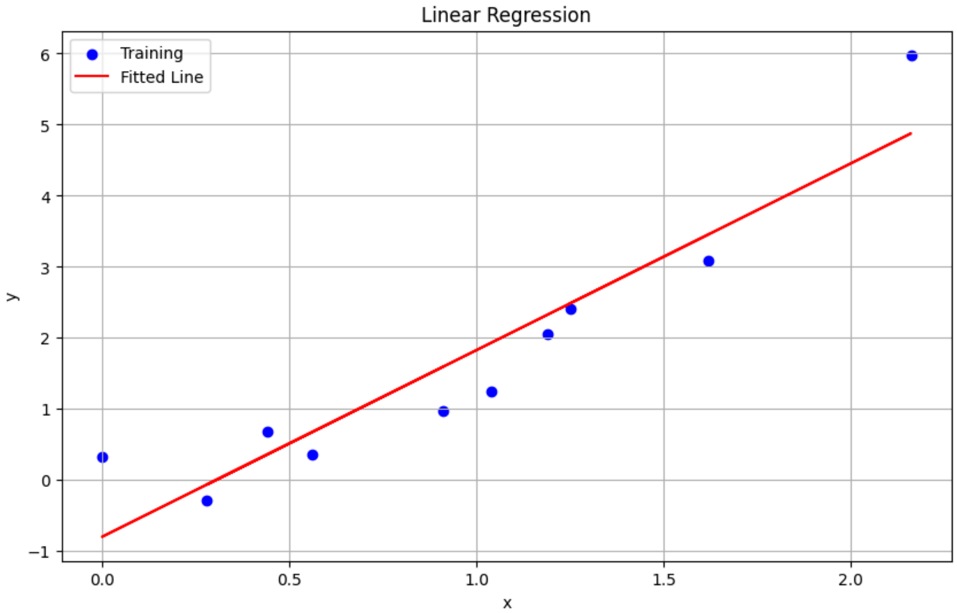


Figure 2: Linear Regression - a

Following the Anayltical Solution providede above, as well as having $X^T = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1.25 & 2.16 & 0.00 & 0.91 & 0.44 & 0.28 & 0.56 & 1.04 & 1.19 & 1.62 \end{pmatrix}$. Then $(X^T X)^{-1} = \begin{pmatrix} 0.333 & -0.247 \\ -0.247 & 0.261 \end{pmatrix}$, $X^T Y = \begin{pmatrix} 16.85 \\ 25.993 \end{pmatrix}$ therefore we can get the result $\Theta^* = \begin{pmatrix} 0.333 & -0.247 \\ -0.247 & 0.261 \end{pmatrix} \begin{pmatrix} 16.85 \\ 25.993 \end{pmatrix} = \begin{pmatrix} -0.7973 & 2.6267 \end{pmatrix}$.

Analytical solution : $\Theta^* = [-0.79725653, 2.62672649]$ hence,

$\therefore f_\Theta(x) = \theta_0 + \theta_1 x = -0.7973 + 2.6267x.$

(b) Suppose we want to increase the model complexity, by considering y as a linear function of both x and x_2. Namely $f_\Theta(x) = \theta_0 + \theta_1 x + \theta_2 x^2$. In this case, calculate the analytic solution of model and plot the smooth curve of the model, together with the scatter plot of data points in training set.

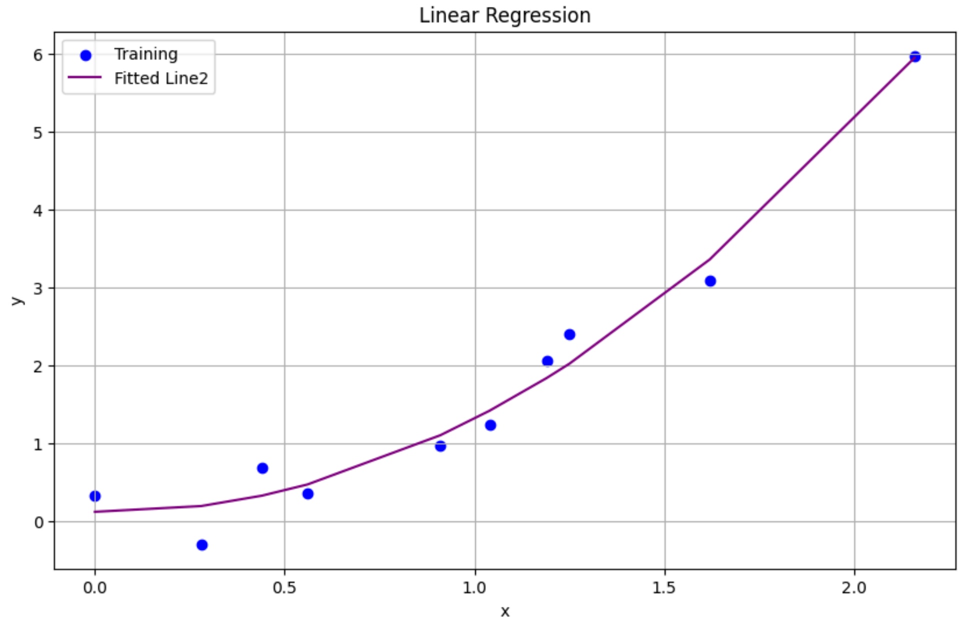


Figure 3: Linear Regression - b

Following the Anayltical Solution providede above, as well as having

$X^T = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1.25 & 2.16 & 0.00 & 0.91 & 0.44 & 0.28 & 0.56 & 1.04 & 1.19 & 1.62 \\ 1.25^2 & 2.16^2 & 0.00^2 & 0.91^2 & 0.44^2 & 0.28^2 & 0.56^2 & 1.04^2 & 1.19^2 & 1.62^2 \end{pmatrix}$. Then $\left(X^T X\right)^{-1} = \begin{pmatrix} 0.609 & -1.064 & 0.388 \\ -1.064 & 2.676 & -1.148 \\ 0.388 & -1.147 & 0.545 \end{pmatrix}$,

$X^T Y = \begin{pmatrix} 16.85 \\ 25.993 \\ 45.071 \end{pmatrix}$ therefore we can get the result

$\Theta^* = \begin{pmatrix} 0.609 & -1.064 & 0.388 \\ -1.064 & 2.676 & -1.148 \\ 0.388 & -1.147 & 0.545 \end{pmatrix} \begin{pmatrix} 16.85 \\ 25.993 \\ 45.071 \end{pmatrix} = \begin{pmatrix} 0.125 & -0.099 & 1.295 \end{pmatrix}$.

Analytical solution : $\Theta^* = [0.12507029, -0.09873162, 1.29523977]$ hence,

$\therefore f_\Theta(x) = \theta_0 + \theta_1 x + \theta_2 x^2 = 0.125 - 0.0987x + 1.295x^2$

(c) Let us further increase the model complexity by assuming y is related to higher-order forms of x, i.e., $f_\Theta(x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$. Again, calculate the analytic solution of the model and plot the curve of the function, together with the data points in training set.

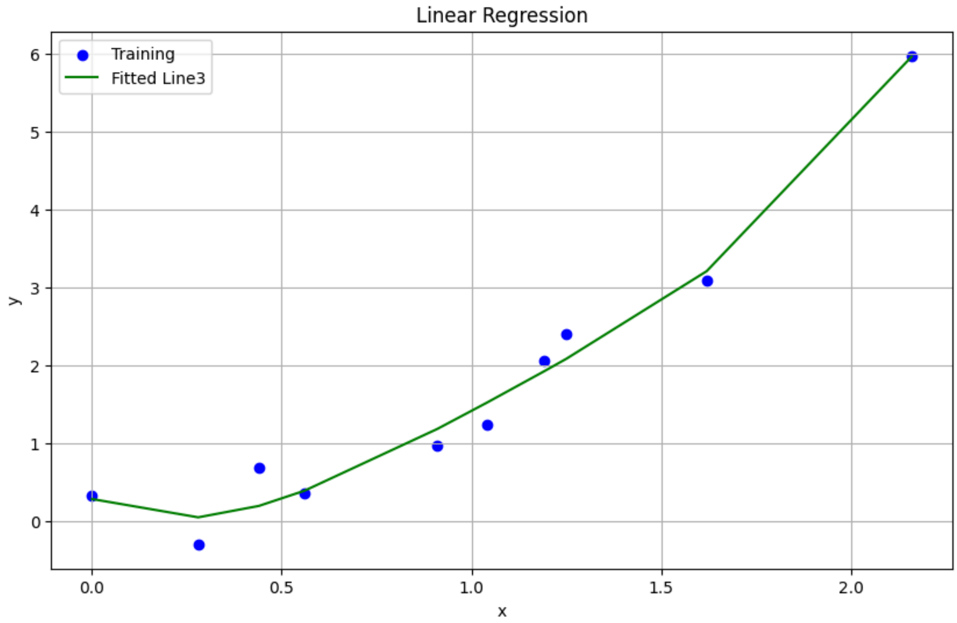


Figure 4: Linear Regression - c

Following the Anayltical Solution providede above, as well as having

$$X^T = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1.25 & 2.16 & 0.00 & 0.91 & 0.44 & 0.28 & 0.56 & 1.04 & 1.19 & 1.62 \\ 1.25^2 & 2.16^2 & 0.00^2 & 0.91^2 & 0.44^2 & 0.28^2 & 0.56^2 & 1.04^2 & 1.19^2 & 1.62^2 \\ 1.25^3 & 2.16^3 & 0.00^3 & 0.91^3 & 0.44^3 & 0.28^3 & 0.56^3 & 1.04^3 & 1.19^3 & 1.62^3 \\ 1.25^4 & 2.16^4 & 0.00^4 & 0.91^4 & 0.44^4 & 0.28^4 & 0.56^4 & 1.04^4 & 1.19^4 & 1.62^4 \end{pmatrix}.$$

$$\text{Then } \left(X^TX\right)^{-1} = \begin{pmatrix} 0.966 & -4.591 & 6.823 & -3.994 & 0.798 \\ -4.591 & 45.679 & -92.031 & 64.152 & -14.313 \\ 6.826 & -92.031 & 207.993 & -154.749 & 35.941 \\ -3.994 & 64.152 & -154.749 & 119.603 & -28.449 \\ 0.798 & -14.313 & 35.941 & -28.449 & 6.871 \end{pmatrix},$$

$$X^TY = \begin{pmatrix} 16.850 \\ 25.993 \\ 45.071 \\ 83.833 \\ 163.656 \end{pmatrix} \text{ therefore we can get the result}$$

$$\Theta^* = \begin{pmatrix} 0.966 & -4.591 & 6.823 & -3.994 & 0.798 \\ -4.591 & 45.679 & -92.031 & 64.152 & -14.313 \\ 6.826 & -92.031 & 207.993 & -154.749 & 35.941 \\ -3.994 & 64.152 & -154.749 & 119.603 & -28.449 \\ 0.798 & -14.313 & 35.941 & -28.449 & 6.871 \end{pmatrix} \begin{pmatrix} 16.850 \\ 25.993 \\ 45.071 \\ 83.833 \\ 163.656 \end{pmatrix} = \begin{pmatrix} 0.289 & -2.337 & 6.319 & -3.705 & 0.854 \end{pmatrix}.$$

Analytical solution : $\Theta^* = [0.28915117, -2.33650044, 6.31881531, -3.7052006, 0.85375681]$ hence,

$\therefore f_\Theta(x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4 = 0.289 - 2.337x + 6.319x^2 - 3.705x^3 + 0.854x^4$

(d) Observe the above three functions, point out which could be faced with underfitting, which could be faced with overfitting, and which one is relatively a good fit? Then, calculate the values of prediction error on the test data to verify your thoughts.

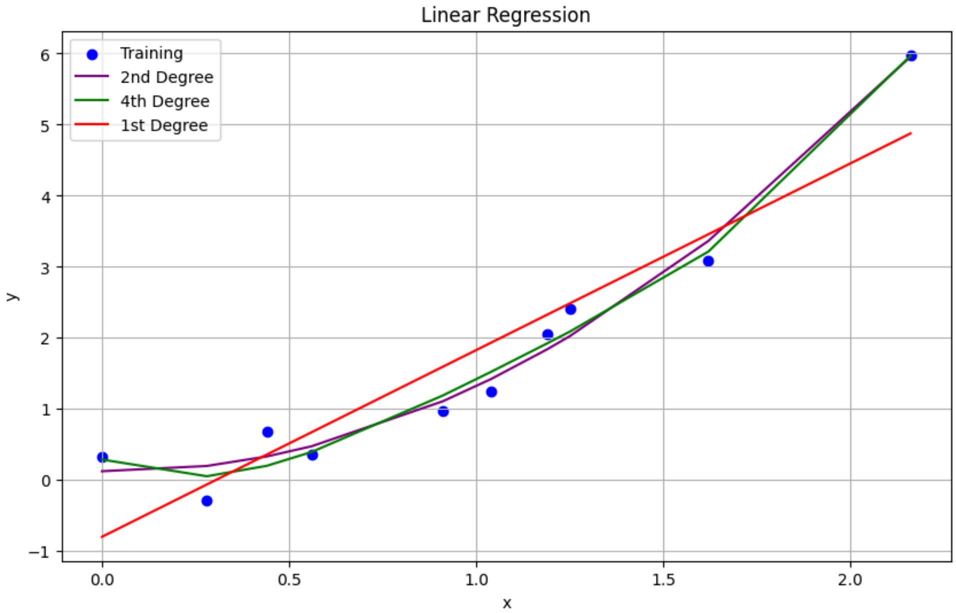


Figure 5: Linear Regression (comprehensive) - d

By observing the three graphs along with the training data, we can assume that the 1st order (figure 2) linear model is an underfit, where as the 4th order linear model (figure 4) is an overfit. Which we can finally assume that the 2nd order linear model (figure 3) is a relatively good fit.

Now let‘s calculate the values of prediction error on the test data. We will identify the R-sqaure score which is determined by $R^2 = 1 - \frac{\sum\left(Y - \widehat{Y}\right)^2}{\sum\left(Y - \overline{Y}\right)^2}$.

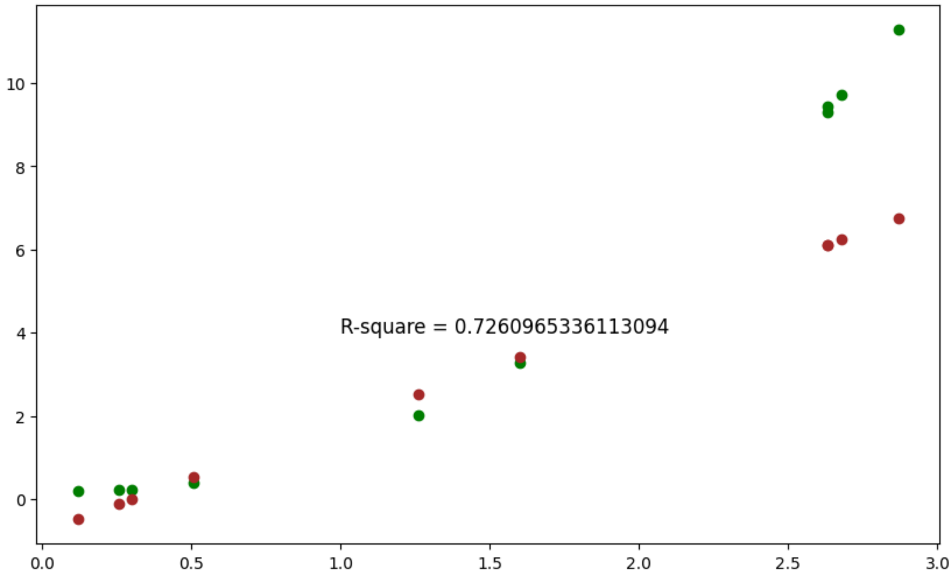


Figure 6: Linear Regression (comprehensive) - d

For the first linear regression model we can calculate that

$$\sum\left(Y - \hat{Y}\right)^2 = 54.775, \sum\left(Y - \overline{Y}\right)^2 = 199.978 \rightarrow R^2 = 1 - \frac{54.775}{199.978} = 1 - 0.274 = 0.726.$$

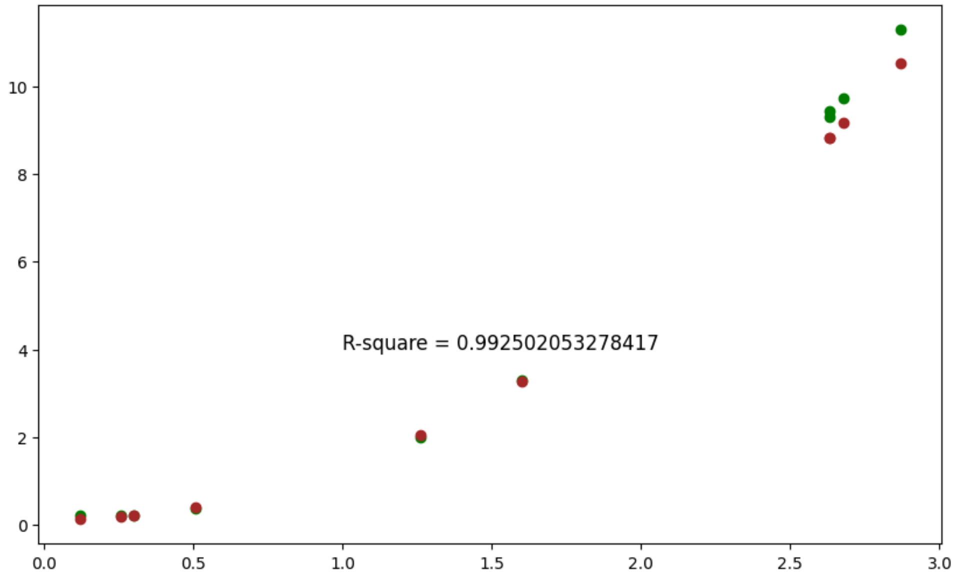


Figure 7: Linear Regression (comprehensive) - d

For the second order linear regression model we can calculate that

$$\sum\left(Y-\hat{Y}\right)^2=1.499,\sum\left(Y-\overline{Y}\right)^2=199.978\rightarrow R^2=1-\frac{1.499}{199.978}=1-0.0075=0.993.$$

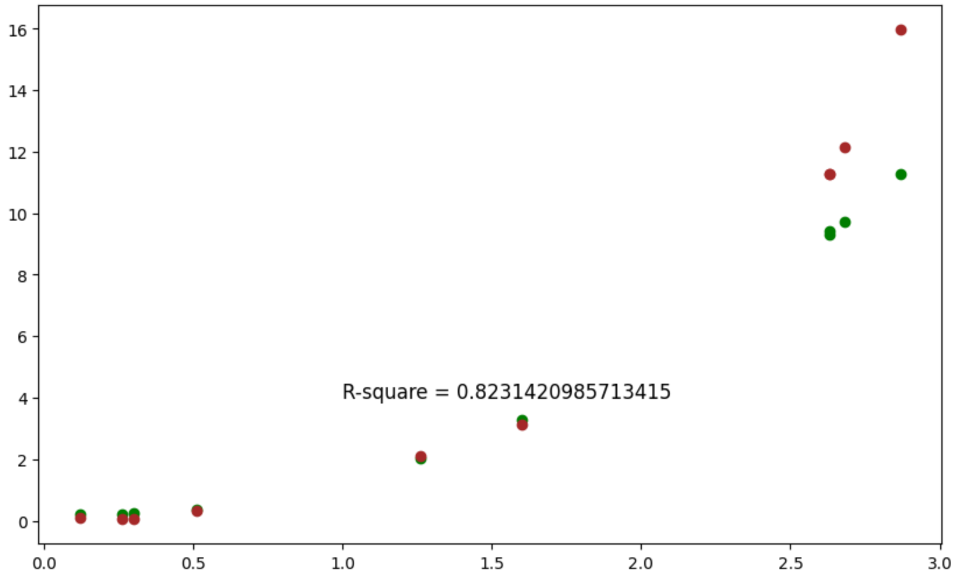


Figure 8: Linear Regression (comprehensive) - d

For the 4th order linear regression model we can calculate that

$$\sum\left(Y-\hat{Y}\right)^2=35.368,\sum\left(Y-\overline{Y}\right)^2=199.978\rightarrow R^2=1-\frac{35.368}{199.978}=1-0.177=0.823.$$

From the above plots we can see that the linear models have a R-square score of approximately 0.726, 0.993 and 0.823 respectively. Due to the fact that an R-square score of near 1 represents a higher prediction accuracy we can conclude that our observation / assumption from above holds.

2) Machine Learning Question

2. a) Question 1 : What is the main objective of a linear regression model?

(A) To classify data points into distinct categories.

(B) To estimate the parameters of a non-linear relationship between variables.

(C) To predict a continuous outcome variable based on one or more predictor variables.

(D) To reduce the dimensionality of the input data.

2. b) Question 2 : What are true for Linear Regression

(A) The assumption of linearity between the dependent variable and the independent variables. In the real world, the data is not always linearly separable

(B) Linear regression is sensitive to outliers

(C) Before applying Linear regression, multicollinearity should be removed because linear regression assumes that there is no relationship among independent variables.