**Student Name & ID: AIML Assignment 2: [80 marks]**

**Regression: Linear, Polynomial, Multiple, Scale, Train/Test**

x = [5,7,8,7,2,17,2,9,4,11,12,9,6]  
y = [90,80,87,88,90,86,89,87,94,78,77,85,86]

Figure 1: the x-axis represents a car’s age, and the y-axis represents its speed.

Using the two vectors in Figure 1 above, write code in Python to:

1. Create a Scatter Plot with the values of x and y [4 marks]

x=[5,7,8,7,2,17,2,9,4,11,12,9,6]

y=[90,80,87,88,90,86,89,87,94,78,77,85,86]

plt.scatter(x,y)

plt.xlabel('Age of car')

plt.ylabel('Speed of car')

plt.show()

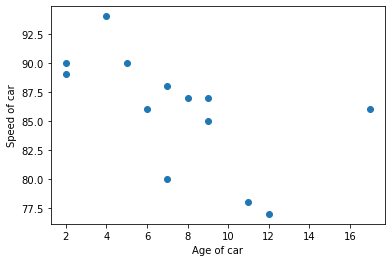


Figure 1 scatter plot for question 1

1. Fit the data to a Linear Regression line, and plot it on the scatter plot [4 marks]

slope,intercept,r,p,std\_err = stats.linregress(x,y)

Reg\_LineY = np.array(x)\*slope + intercept

plt.scatter(x,y)

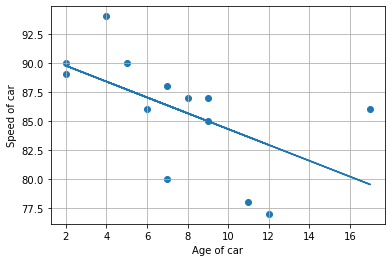
plt.plot(x,Reg\_LineY)

plt.xlabel('Age of car')

plt.ylabel('Speed of car')

plt.grid()

plt.show()



1. Estimate the coefficient of correlation, r, and explain what it means [4 marks]

print(r)



The coefficient of correlation r ranges from 1 to -1. It is used as a measurement of the relationship between 2 variables usually x and y. Any value of r that is greater than 1 or -1 represents an error. Values of r that are close to 1 and -1 represent better relationship between variables. If they are exactly 1 and -1, they represent perfectly positive and negative relationships between the variables. Values that are closer to 0 represent a poor relationship between 2 variables.

In this case, an r value of -0.5753 represents an adequate or average negative relationship between the car year (x) and the speed of the car(y).

1. Predict the speed of a 13 year old car using the linear reg. model [4 marks]

Car\_Age=13

Speed\_Predict= Car\_Age \*slope + intercept

print(Speed\_Predict)



1. Fit the data to a third degree polynomial and plot it on the scatter plot[4 marks]

n=3

mymodel=np.poly1d(np.polyfit(x,y,n))

myline=np.linspace(1,17,100)

plt.scatter(x,y)

plt.plot(myline,mymodel(myline))

plt.xlabel('Age of car')

plt.ylabel('Speed of car')

plt.grid()

plt.show()

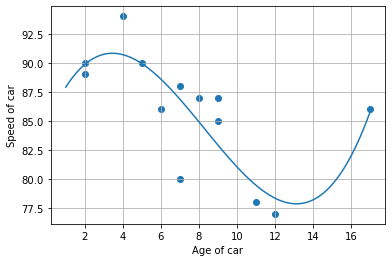


Figure 2 3rd degree polynomial plot for Q5

1. Predict the speed of a 13 year old car using the 3rd deg. polynomial [4 marks]

speed=mymodel(13)

print(speed)



As seen from the image above the predicted speed of a 13 year old car using the 3rd degree polynomial is 77.876

1. Estimate the r-squared value and explain what it means [4 marks]

print(r2\_score(y,mymodel(x)))



As seen from the image above, the estimated r-squared value is 0.6718. The r squared value ,also known as the coefficient of determination, is used to determine the relationship between x axis values and the y axis values. The r squared value ranges from 0 to 1 where a value of ‘0’ indicates that there is no relationship between the x and y axis values and a ‘1’ indicates that the values of the x and y axis are 100% related. If there is no relationship the polynomial regression cannot be used for prediction. In this case , the r-squared value was estimated to be 0.6718. This indicates that the age of the car (x axis values) are 67.18% related to the speed of the car (y axis values).

1. Explain any differences in values obtained in (4) and (6) [4 marks]

The predicted speed of a 13 year old car using the linear regression model is : 82.256 with a r value of -0.57535

The predicted speed of a 13 year old car using the polynomial regression model is : 77.876 with a r-squared value of 0.6718.

From the values stated above , it can be seen that the predicted speed from the linear model was higher than that of the polynomial model. There are several reasons for this difference.

* The linear model best fit line has several datapoints that are very far from it, [(7,80) , (11,78), (12,77) ,(17,86)]. Since these points are far from the best fit line, it can be said that the line is not a perfect fit which means that the prediction can be quite higher or lower compared to the actual data points. Compared to the polynomial model line, the are only 2 points that are far from the best fit line [(4,94) , (5,80)] , this means that the best fit line can provide a better prediction that is closer to the actual data points since it has less outliers.
* This first point can also be illustrated by looking at the r and r-squared values. The r-squared value of the polynomial model is closer to the perfect relationship indicator (r-squared value of ‘1’) than the linear model r value (r value of ‘-1’).This fact can then indicate that the polynomial model has a better/stronger relationship between axes than the linear model which means that the predicted value will be closer to actual data points. This results in the prediction of the polynomial model being smaller than the linear model since it is more accurate.

Use the attached excel file data1.csv to perform the following:

1. Read in the data1 file into your workspace and store it as a data frame[4 marks]

dir\_path = r"C:\Users\Admin\Desktop\compEng\Year 3\Semester2\AIML\AIML\Assignment1"

file\_name = "data1.csv"

file\_path = os.path.join(dir\_path, file\_name)

# reading in the file as an object

df = pd.read\_csv(file\_path)

1. Extract the weight and volume fields and store it in a variable, X [2 marks]

X = df[["Weight","Volume"]]

1. Extract the CO2 field and store it as variable, y [2 marks]

y = df[["CO2"]]

1. Fit a Linear Regression model to variables X and y [4 marks]

linReg = linear\_model.LinearRegression()

linReg.fit(X, y)

1. Predict value of CO2 of the Volvo XC70 with weight of 1746 kg & volume of 2000cm3 , and explain why this value is different from 117, as given in the data file. [6 marks]

predictedCO2 = linReg.predict([[1746, 2000]])



The predicted CO2 of the Volvo XC70 with weight of 1746 kg & volume of 2000cm3 is : 108.7182 grams.

This is smaller than the value in the data1.csv file which states that the CO2 value is 117g. This discrepancy can be due several factors.

* The varying values of weight and volumes may cause the model to not perfectly fit the data which can result in a different value from the actual datapoint
* The multiple regression model uses several assumptions such as there being a linear relationship between the x and y values and that the x values are not too correlated to each other. Both these assumptions can results in differences in the predicted value.
* Every model has some error value called the residual value, E, which is the difference between the predicted value and the actual value, to compensate for it not being able to perfectly predict values.

1. Find the reg. coefficient between X and y and explain its meaning [4 marks]

print(linReg.coef\_)



The regression coefficients for this model is :

Weight = 0.00766976

Volume = 0.00818285

This means that for every 1 kg of weight increase, the CO2 output increases by 0.00766976g and for every 1 of volume increase the CO2 output increases by & 0.00818285g.

1. Find the R2 (R-squared) score and explain its meaning [4 marks]

print(linReg.score(X,y))

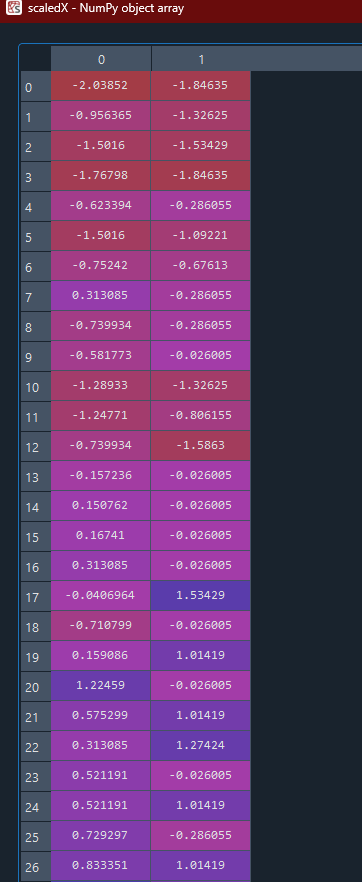
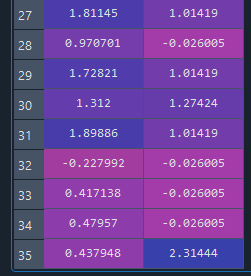


The R2 score is : 0.4090. The R2 score is used to measure the relationship between the xaxis and the y axis values. It ranges from 0 to 1, where ‘0’ means no relationship and a ‘1’ indicates a perfect relationship. Since the R2 value obtained in this case is 0.4090 , it means that there is a relationship between the x and y axis values since it is greater than ‘0’ however the relationship is mediocre since it is not particularly close to ‘1’.

1. Scale the variables in X using the standardization method [4 marks]

scale = StandardScaler()

scaledX = scale.fit\_transform(X)

1. Repeat question (13) using this scaled X variable [2 marks]

linRegScaled = linear\_model.LinearRegression()

linRegScaled.fit(scaledX, y)

scaledValue = scale.transform([[1746, 2000]])

scaledPredictedCO2 = linRegScaled.predict([scaledValue[0]])



1. Explain any differences in values in ques: (13) & (17) [2 marks]

The predicted CO2 of the Volvo XC70 with weight of 1746 kg & volume of 2000cm3 in question 13 is : 108.7182 grams.

The predicted CO2 of the Volvo XC70 with weight of 1746 kg & volume of 2000cm3 in question 17 is : 108.7182 grams.

Both Predicted CO2 values are the same thus there is no difference. There is no difference between the values since the scaled values preserve the relationship between the x axis values and the y axis values.

1. Explain what is the train/test method and why is it important [4 marks]

The train/ test method is used to determine the accuracy of a model. This method as its name suggests involves training and testing. The data set is split into a training set and a testing set, where 80% of the data is used for training and the remaining 20% of data is used for the testing. Training the model means creating the model and testing the model means testing how accurate the model is.

1. Apply the train/test method as follows: i) use the first 30 values of variables X and y to train a Linear Regression model, ii) repeat question (13), (14), (15) and iii) explain any differences between the values obtained in here and in question (13) for the value of the CO2 (calculate % differences between predicted and actual value for your discussion) for the Volvo XC70, the reg. coefficient and R-squared values. [10 marks]

Please type out your solutions in this word document just after the respective question (so that I can see the marks allocated when I am marking), then convert it to a pdf (file 🡺 save as 🡺 type pdf, make sure you have Adobe Acrobat Reader installed) submit your solutions as a pdf file on canvas