

# Batch:A2 Roll No.:16010421075 Experiment No.: 3

**Aim:** Predict missing data using regression modelling.

**Resources needed:** Any programming language, any data source (RDBMS/Excel/CSV)

# Theory:

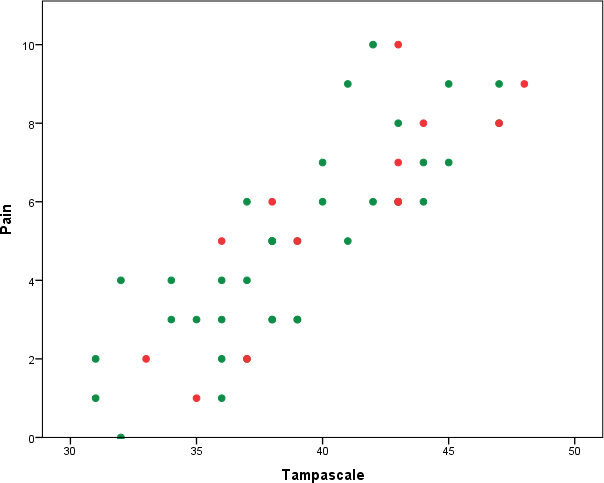
Missing data (or missing values) is defined as the data value that is not stored for a variable in the observation of interest. The problem of missing data is relatively common in data set and can have a significant effect on the conclusions that can be drawn from the data. There are various techniques proposed for handling missing values like deletion of records/attributes, filling with a random value or using some measures of central tendency, imputation using regression etc. Regression imputation is guessing missing variables using regression if we know there is a correlation between the missing value and other variables. Scatterplots can be used to identify correlation between variables.

Figure 1: A scatter plot showing correlation between attributes pain and tampascale.

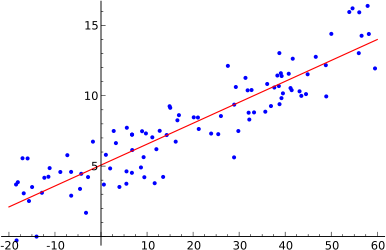
Once correlation is identified either linear regression or multiple regression can be used for imputation. Linear regression involves finding the “best” line as shown in fig. 1 to fit two attributes (or variables) so that one attribute can be used to predict the other.

Figure 2: Example of simple linear regression, which has one independent variable

*Multiple linear regression* is an extension of linear regression, where more than two attributes are involved and the data are fit to a multidimensional surface.

Prediction is predicting continuous or ordered values for a given input i.e. Numeric prediction, for example, predicting salary of employee with 10 years of experience.

# Simple Linear Regression:

Straight line regression analysis involves a responsible variable *y* and a single predictor variable *x*. by modelling *y* as a linear function of *x* as given in equation 1

*y=w0 + w1\*x* (1)

where *w0* and *w1* are Regression co-efficient.

*w0 = Y-intercept*

*w1 = Slope of the line*

Calculate *w0* and *w1* by method of least squares, which estimates best fitting straight line. Let *D* be a training set,

[ *D* ] = *{ (x1,y1),(x2,y2),(x3,y3),… ,( xn, yn)}*

Regression co-efficient,

𝑤 =

1

|𝐷|

∑

𝑖=1

(𝑖−𝑥)(𝑖−𝑦)

………………………………………………………………..(2)

𝑤 = 𝑦 − 𝑤 𝑥…………………………………………………..…………………….(3)

0 1

Where 𝑥 is the mean value of *x1, x2, x3,….xn*.

And 𝑦 is the mean value of *y1, y2, y3, y4,…yn.*

# Multiple Linear Regression:

Multiple linear regression is used to explain the relationship between one continuous dependent variable and two or more independent variables. The independent variables can be continuous or categorical.

𝑦 = 𝑏 + 𝑏 𝑥 +𝑏 𝑥 +…+ 𝑏 𝑥 +ε

𝑖 0 1 𝑖1 2 𝑖2 𝑛

Where , for i=1 to n observations:

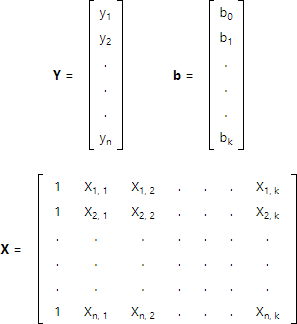
𝑦𝑖 = dependent variable

𝑥𝑖1,𝑥𝑖2, = predictor variables

𝑏 , 𝑏 𝑏 = Regression coefficients

1 2, 3.......

ε = Model’s error term



To handle the complications of multiple regression, we will use matrix algebra. The least squares normal equations can be expressed as: **Y=Xb** Multiply both sides with XT

**XTY** = **XTXb** or **XTXb** = **XTY**

Here, matrix **XT** is the transpose of matrix **X**. To solve for regression coefficients, simply pre-multiply by the inverse of **XTX**:

**(XTX)-1XTXb**=**(XTX)-1XTY** since**(XTX)-1XTX**=**I**, the identity matrix, we get slope b

as,

**b**=**(XTX)-1XTY**

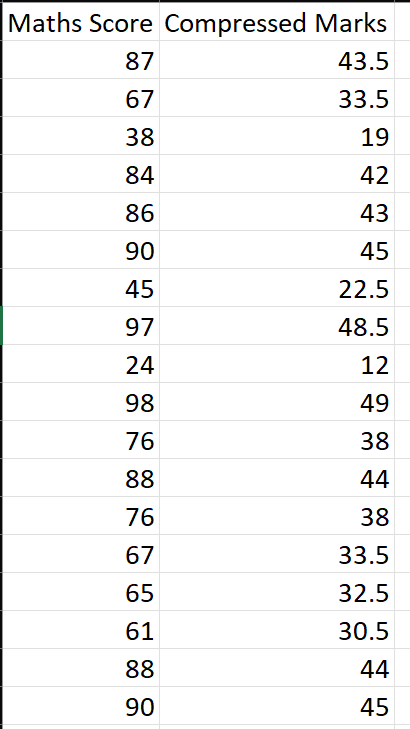
# Activity :

1. Identify attributes suitable for applying Linear regression. Construct a linear regression model for your dataset and predict the missing values in your data set. Evaluate the accuracy of prediction.(use of pre defined built in package for prediction is not expected)
2. Identify attributes suitable for applying Multiple Linear regression. Construct a linear regression model for your dataset and predict the missing values in your data set. Evaluate the accuracy of prediction.(use of pre defined built in package for prediction is not expected)

# Results: Students must submit the programs along with the output in a separate

**\*.doc/.docx/.txt file along with post lab questions.**

**Dataset**

****

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

df = pd.read\_csv("dataset.csv")

c1 = df["Maths Score"]

c2 = df["Compressed Marks"]

correlation = c1.corr(c2)

def relation():

    if correlation == 1.0:

        print("Relation: Linear")

    else:

        print("Relation: Not Linear")

def dfmean1():

    maths\_marks = df["Maths Score"]

    total1 = maths\_marks.sum()

    count1 = len(maths\_marks.axes[0])

    mean1 = total1/count1

    print("x(mean) =", mean1)

def dfmean2():

    compressed\_marks = df["Compressed Marks"]

    total2 = compressed\_marks.sum()

    count2 = len(compressed\_marks.axes[0])

    mean2 = total2/count2

    print("y(mean) =", mean2)

print("\n")

relation()

dfmean1()

dfmean2()

def estimate\_coef(x, y):

    n = np.size(x)

    m\_x = np.mean(x)

    m\_y = np.mean(y)

    SS\_xy = np.sum(y\*x) - n\*m\_y\*m\_x

    SS\_xx = np.sum(x\*x) - n\*m\_x\*m\_x

    global W\_0

    global W\_1

    W\_1 = SS\_xy / SS\_xx

    W\_0 = m\_y - W\_1\*m\_x

    return (W\_0, W\_1)

def plot\_regression\_line(x, y, b):

    plt.scatter(x, y, color="m",

    marker="o", s=30)

    y\_pred = b[0] + b[1]\*x

    plt.plot(x, y\_pred, color="g")

    plt.xlabel('x')

    plt.ylabel('y')

    plt.show()

def main():

    x = df["Maths Score"]

    y = df["Compressed Marks"]

    b = estimate\_coef(x, y)

    print("Estimated coefficients:\nW\_0 = {} \

    \nb\_1 = {}".format(b[0], b[1]))

    plot\_regression\_line(x, y, b)

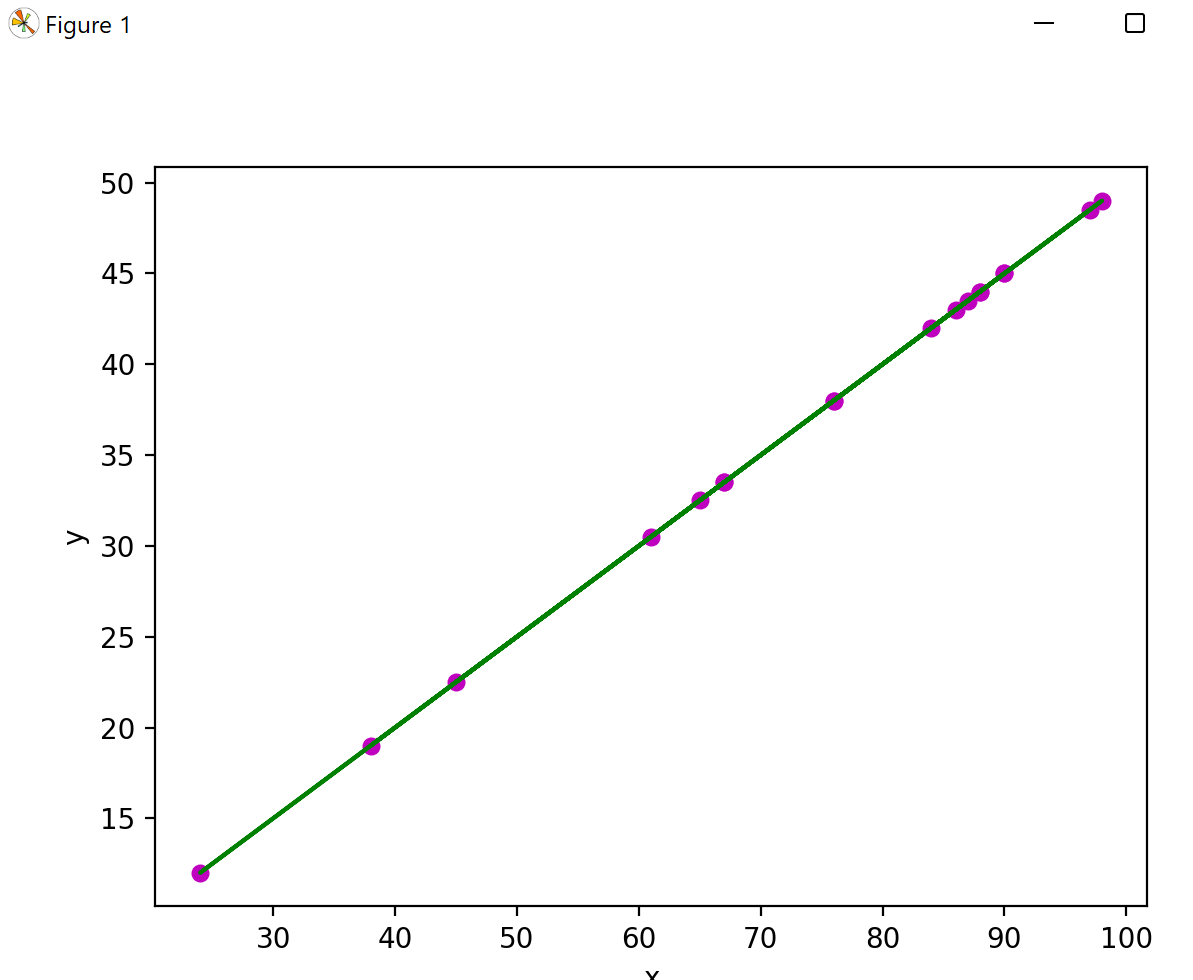
if \_\_name\_\_ == "\_\_main\_\_":

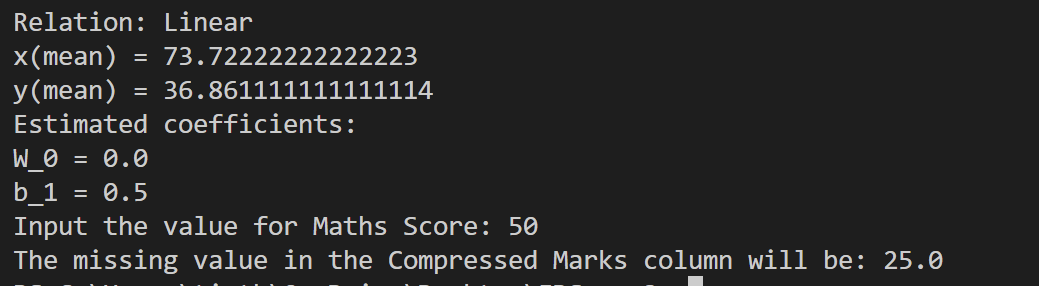
    main()

y = int(input("Input the value for Maths Score: "))

x = y\*W\_1 +W\_0

print("The missing value in the Compressed Marks column will be:", x)

****

****

# Questions:

* 1. How will you choose between linear regression and non-linear regression?

Ans) Linear regression is easier to use, simpler to interpret, and you obtain more statistics that help you assess the model. While linear regression can model curves, it is relatively restricted in the shapes of the curves that it can fit. Sometimes it can’t fit the specific curve in your data. Nonlinear regression can fit many more types of curves, but it can require more effort both to find the best fit and to interpret the role of the independent variables. Additionally, R-squared is not valid for nonlinear regression, and it is impossible to calculate p-values for the parameter estimates.

* 1. Explain the nature or characteristics of a dataset where we can apply regression imputation.

Ans) Regression imputation fits a statistical model on a variable with missing values. Predictions of this regression model are used to substitute the missing values in this variable. Mean, median or mode imputation only look at the distribution of the values of the variable with missing entries. If we know there is a correlation between the missing value and other variables, we can often get better guesses by regressing the missing variable on other variables.

# Outcomes: CO2: Comprehend descriptive and proximity measures of data

**Conclusion: To be written by students based on their understanding of the experiment**

Linear regression program was successfully implemented.

# Grade: AA / AB / BB / BC / CC / CD /DD

Signature of faculty in-charge with date

# References:

Books/ Journals/ Websites:

1. Han, Kamber, "Data Mining Concepts and Techniques", Morgan Kaufmann 3nd Edition