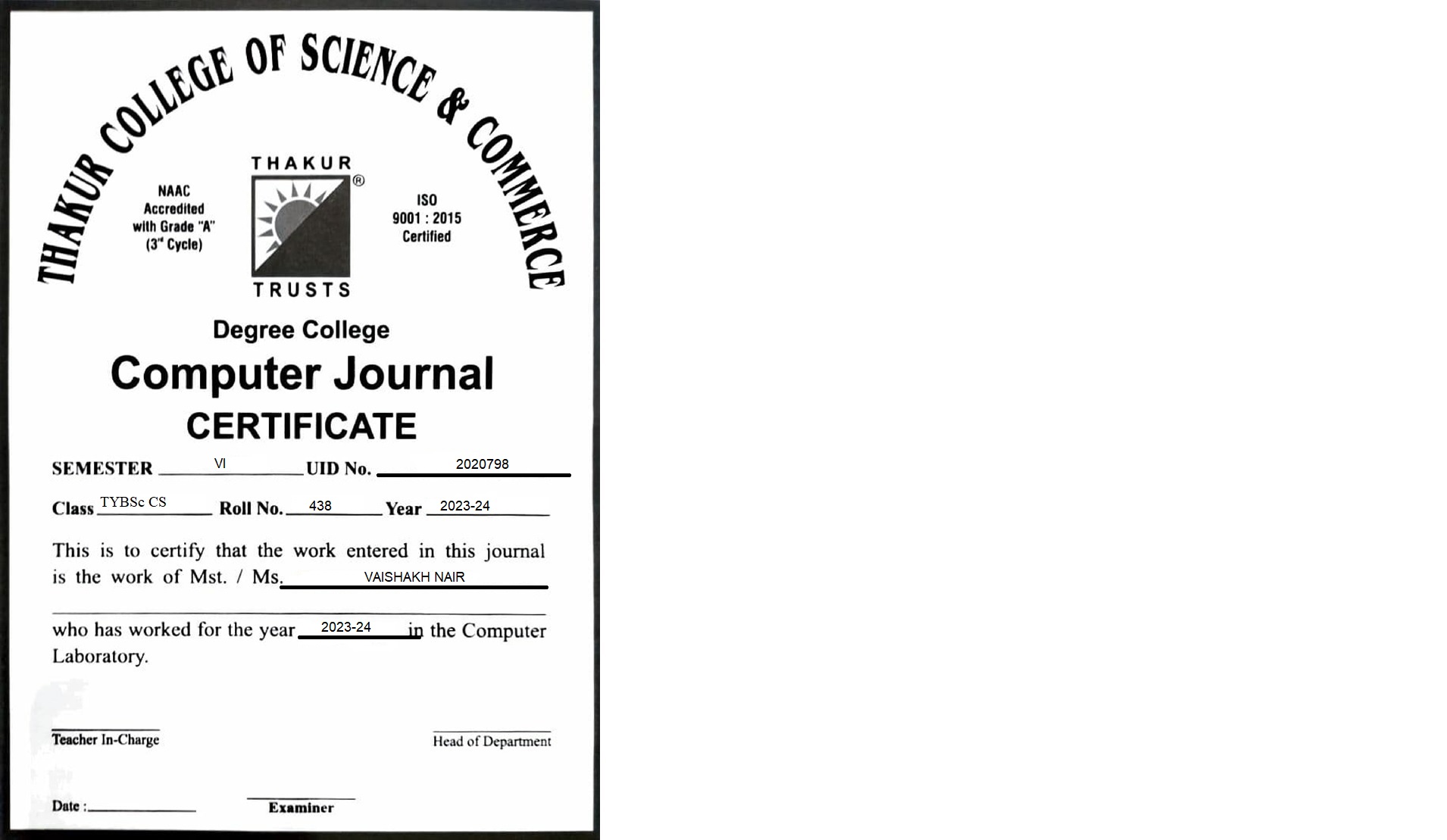
NAME – VAISHAKH NAIR

CLASS – TYBSC CS

SUB – DATA SCIENCE

ROLL NO – 438



**AIM**: Demo of Simple/Multiple Linear Regression

**THEORY**:

1. Simple Linear Regression:

Simple linear regression is used to estimate the relationship between two quantitative variables. You can

use simple linear regression when you want to know:

• How strong the relationship is between two variables (e.g., the relationship between rainfall and

soil erosion).

• The value of the dependent variable at a certain value of the independent variable (e.g., the

amount of soil erosion at a certain level of rainfall).

Regression models describe the relationship between variables by fitting a line to the observed data.

Linear regression models use a straight line, while logistic and nonlinear regression models use a curved

line. Regression allows you to estimate how a dependent variable changes as the independent variable(s)

change.

For Eg:

You are a social researcher interested in the relationship between income and happiness. You survey 500

people whose incomes range from 15k to 75k and ask them to rank their happiness on a scale from 1 to

10.

Your independent variable (income) and dependent variable (happiness) are both quantitative, so you can

do a regression analysis to see if there is a linear relationship between them.

2. Multiple Linear Regression:

Multiple linear regression is used to estimate the relationship between two or more independent

variables and one dependent variable. You can use multiple linear regression when you want to know:

• How strong the relationship is between two or more independent variables and one dependent

variable (e.g. how rainfall, temperature, and amount of fertilizer added affect crop growth).

• The value of the dependent variable at a certain value of the independent variables (e.g. the

expected yield of a crop at certain levels of rainfall, temperature, and fertilizer addition).

For Eg:

You are a public health researcher interested in social factors that influence heart disease. You survey 500

towns and gather data on the percentage of people in each town who smoke, the percentage of people in

each town who bike to work, and the percentage of people in each town who have heart disease.

Because you have two independent variables and one dependent variable, and all your variables are

quantitative, you can use multiple linear regression to analyze the relationship between them.

CODE AND OUTPUT:

import numpy as np

import matplotlib.pyplot as plt

def estimate\_coef(x, y):

# number of observations/points

n = np.size(x)

# mean of x and y vector

m\_x = np.mean(x)

m\_y = np.mean(y)

# calculating cross-deviation and deviation about x

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

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

# calculating regression coefficients

b\_1 = SS\_xy / SS\_xx

b\_0 = m\_y - b\_1\*m\_x

return (b\_0, b\_1)

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

# plotting the actual points as scatter plot

plt.scatter(x, y, color = "m",

marker = "o", s = 30)

# predicted response vector

y\_pred = b[0] + b[1]\*x

# plotting the regression line

plt.plot(x, y\_pred, color ="g")

# putting labels

plt.xlabel('x')

plt.ylabel('y')

# function to show plot

plt.show()

def main():

# observations / data

x = np.array([0, 1, 2, 3, 4, 5, 6, 7, 8, 9])

y = np.array([1, 3, 2, 5, 7, 8, 8, 9, 10, 12])

# estimating coefficients

b = estimate\_coef(x, y)

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

\nb\_1 = {}".format(b[0], b[1]))

# plotting regression line

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

main()

OUTPUT:

