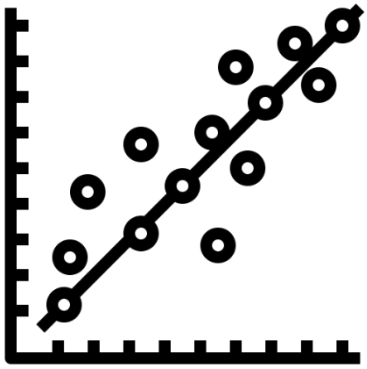
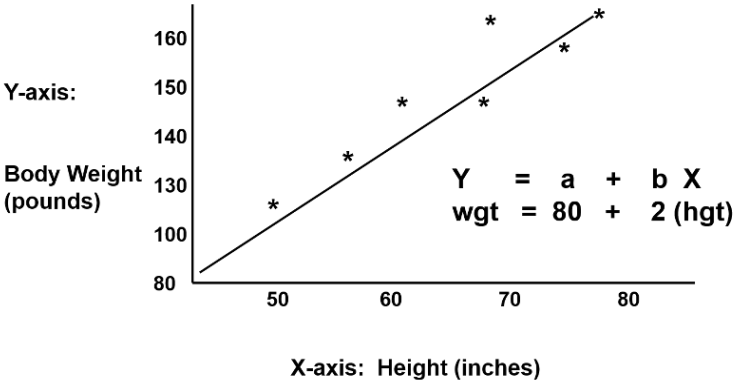
Dr. A. Q. Khan Institute of Computer Sciences & Information Technology, (KICSIT)
Department of Computer Engineering

Artificial Intelligence Lab (03)

Implementation of Linear Regression Supervised Machine Learning Algorithm in Python from Scratch

What is Linear Regression?





- ***Linear regression is a statistical method used to establish a relationship between a dependent variable and one or more independent variables***.
- It involves fitting a straight line to the data points to make predictions.

Dependent Variable: The dependent variable is the one you are trying to predict or explain in regression analysis; it is the outcome/output or response variable or target variable.
Independent Variable: The independent variable is the one used to make predictions or explain the variation in the dependent variable; it is the predictor or explanatory/Input variable.

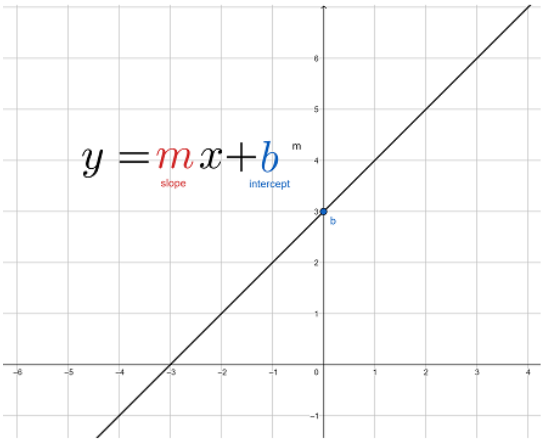
Where it can used?

- In healthcare, linear regression can be used to model the relationship between patient characteristics and health outcomes.
- In marketing, it can be used to determine the impact of advertising or promotional campaigns on sales.
- In sports, it can be used to analyze the relationship between a team's performance and various factors such as player statistics, weather conditions, or game strategies.
- In environmental science, it can be used to study the relationship between pollutants and their effects on the ecosystem.

Formula

$$y = mx + c$$

y = Output / Dependent Variable.
m = Slope.
x = Input / Independent Variable.
c = Constant / y-intercept.



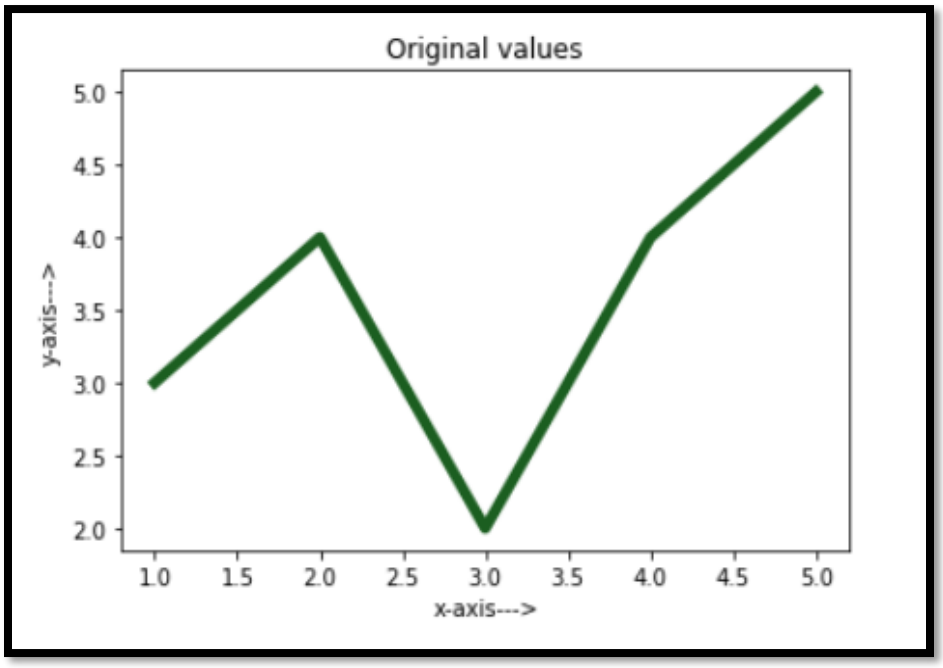
Lab Instructor: Engr. Zia Ur Rehman

- The **intercept** is the value where the regression line crosses the y-axis. It represents the predicted value of the dependent variable when all independent variables are equal to zero.
- The **intercept is called an intercept** because it represents the point where the regression line intercepts the vertical y-axis on a graph. It is also sometimes referred to as the "y-intercept" because it is the point where the regression line intersects the y-axis, which is the vertical axis in a two-dimensional coordinate system.
- The **slope** represents the change in the dependent variable (y) for a one-unit change in the independent variable (x). It is the rate at which the dependent variable changes with respect to the independent variable. The slope is also sometimes referred to as the "**regression coefficient**" or "**beta coefficient**."

Example 01

We have some values of x or input or independent variable given and on behalf of these values we also have values of y or output or dependent variable, and now we want to predict values of y on new values of x.

x	***y (Actual Values)***
1	3
2	4
3	2
4	4
5	5
6	?
7	?
8	?



- To perform a linear regression, we can use the least squares method to find the line of best fit that minimizes the sum of the squared errors between the predicted values and the actual values.
- In a linear regression analysis, the goal is to find a line that best fits the observed data. The line is characterized by an intercept and a slope. The least squares method involves finding the values of the intercept and slope that minimize the sum of the squared differences between the predicted values of the dependent variable (y) and the actual values of the dependent variable (y) for a given set of independent variable (x) values.

Dr. A. Q. Khan Institute of Computer Sciences & Information Technology, (KICSIT)
Department of Computer Engineering

- The equation for a simple linear regression model is:

$$y = mx + c$$

- where y is the dependent variable, x is the independent variable, c is the y-intercept, and m is the slope of the line.
- Using the dataset above, we can calculate the **slope** and **intercept** of the line of best fit as follows:

Formula for Slope *m*?

$$m = \frac{cov(x,y)}{var(x)} = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\Sigma(x-\bar{x})^2}$$

- ***cov*: covariance, *var*: variance , $\bar{x}$ = mean of x, $\bar{y}$ = mean of y**

Covariance: Covariance is a measure of how much two variables are linearly related to each other. In this case, it measures how much the values of x and y vary together.

When the covariance is positive, it indicates that as x increases, y tends to increase as well. When the covariance is negative, it indicates that as x increases, y tends to decrease. If the covariance is zero, it indicates that there is no linear relationship between x and y.

Variance: Variance is a statistical measure of how much the values in a dataset vary from the mean (average) value. It measures the spread of the data around the mean. A high variance indicates that the data points are widely spread out, while a low variance indicates that the data points are closely clustered around the mean.

Finding *m*(*slope*)?

x	y	$(x-\bar{x})$	$(y-\bar{y})$	$(x-\bar{x})^2$	$(x-\bar{x})(y-\bar{y})$
1	3	-2	-0.6	4	1.2
2	4	-1	0.4	1	-0.4
3	2	0	-1.6	0	0
4	4	1	0.4	1	0.4
5	5	2	1.4	4	2.8
mean $\bar{x} = 3$	**mean** $\bar{y} = 3.6$			$\Sigma = 10$	$\Sigma = 4$

$$m = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\Sigma(x-\bar{x})^2}$$

$$m = \frac{4}{10} = 0.4$$

Finding c(*y-intercept*)?

$$c = \bar{y} - (m)(\bar{x}) = 3.6 - (0.4)(3) = 2.4$$

Final Equation after putting value of m and c: This is the linear regression model equation for our scenario to predict the future value of y by putting the value of x. The model is trained by given dataset of x and y.

$$y = 0.4x + 2.4$$

Now predict the value of y when x = 6, 7, 8

$$for\ x = 6, \boldsymbol{y = 0.4(6) + 2.4 = 4.8}$$

$$for\ x = 7, \boldsymbol{y = 0.4(7) + 2.4 = 5.2}$$

$$for\ x = 8, \boldsymbol{y = 0.4(6) + 2.4 = 5.6}$$

Model Evaluation

R Square

R-squared value is a statistical measure of how close the data are to the fitted the regression line. It is also known as coefficient of determination, or the coefficient of multiple determination. If value of R^2 = 1 its mean no error occur 100 % line is fitted. If value of R^2 = 0 its mean error occur 100 % line is not fitted.

$$\boldsymbol{R^2 = \frac{\sum(y_p - \bar{y})^2}{\Sigma(y - \bar{y})^2}}$$

y_P *is Predicted values of* y

x	y	$(y - \bar{y})$	$(y - \bar{y})^2$	$y_{P=0.4x+2.4}$	$(y_p - \bar{y})$	$(y_p - \bar{y})^2$	$R^2 = \frac{\sum(y_p - \bar{y})^2}{\Sigma(y - \bar{y})^2}$
1	3	-0.6	0.36	2.8	-0.8	0.64	
2	4	0.4	0.16	3.2	-0.4	0.16	
3	2	-1.6	2.56	3.6	0	0	
4	4	0.4	0.16	4.0	0.4	0.16	
5	5	1.4	1.96	4.4	0.8	0.64	$\boldsymbol{R^2 = \frac{1.6}{5.2} = 0.30}$
	mean $\bar{y} = 3.6$		$\Sigma = 5.2$			$\Sigma = 1.6$	**Accuracy = 30 %**

Dr. A. Q. Khan Institute of Computer Sciences & Information Technology, (KICSIT)
Department of Computer Engineering

Lab Task/Lab Report

Assigned Date:

1. **Lab Task 01**: Try to implement the Example 01 using Python Programming Language. → link (**Click here**).

[CLO-01, PLO-02, P-3(Guided Response), Rubric (Coding)]

Marks	1	2	3	4
Coding	The code is not as per guidelines and requirements are not met	Some section of code is correct	Most section of code is correct and understands it well	The code is properly written, and have good understanding about it

2. **Lab Task 02**: Let's say you want to **predict a student's final exam score based on the number of hours they study**. In this case, the number of hours studied is the independent variable, also known as the predictor variable, and the final exam score is the dependent variable, also known as the response variable. Here's a sample dataset of 6 students with their corresponding hours studied and final exam scores:

   Make machine learning model for this scenario and implement in python

Hours Studied	***Final Exam Score***
2	60
3	70
4	80
5	85
6	90
7	95

x	y	$(x-\bar{x})$	$(y-\bar{y})$	$(x-\bar{x})^2$	$(x-\bar{x})(y-\bar{y})$	$m = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\Sigma(x-\bar{x})^2}$ $m = ?$
2	60					
3	70					
4	80					
5	85					
6	90					
7	95					
mean $\bar{x} =$	**mean** $\bar{y} =$			$\Sigma =$	$\Sigma =$	

Lab Instructor: Engr. Zia Ur Rehman

Dr. A. Q. Khan Institute of Computer Sciences & Information Technology, (KICSIT)
Department of Computer Engineering

x	y	$(y-\bar{y})$	$(y-\bar{y})^2$	$y_{P=}$	$(y_p-\bar{y})$	$(y_p-\bar{y})^2$	$R^2=\frac{\sum(y_p-\bar{y})^2}{\Sigma(y-\bar{y})^2}$
2	60						
3	70						
4	80						$R^2=?$
5	85						
6	90						**Accuracy = ?**
7	95						
	mean $\bar{y}=$		$\Sigma=$			$\Sigma=$	

[CLO-02, PLO-03, P-4(Mechanism), Rubric (Model Implementation)]

Marks	**1**	**2**	**3**	**4**
Model Implementation	The model is not implemented as per guidelines and requirements are not met	Some section of model is correctly implemented	Most section of model is correctly implemented and understands it well	The model is properly implemented, and have good understanding about it

Dr. A. Q. Khan Institute of Computer Sciences & Information Technology, (KICSIT)
Department of Computer Engineering

Artificial Intelligence Lab
All Rubrics of Microprocessor & Interfacing Lab

CLO 1

Marks	1	2	3	4
Coding	The code is not as per guidelines and requirements are not met	Some section of code is correct	Most section of code is correct and understands it well	The code is properly written, and have good understanding about it

CLO 2

Marks	1	2	3	4
Model Implementation	The model is not implemented as per guidelines and requirements are not met	Some section of model is correctly implemented	Most section of model is correctly implemented and understands it well	The model is properly implemented, and have good understanding about it

CLO 3

Marks	1	2	3	4
Data Pre-processing	The data is not pre-processed as per guidelines and requirements are not met	Some section of data pre-processing is correct	Most section of data pre-processing is correct and understands it well	The data pre-processing is done properly, and have good understanding about it

CLO 4

Marks	1	2	3	4
Team Work	Rarely listens to, shares with, and supports the efforts of others. Often is not a good team member.	Often listen to, shares with and supports the efforts of others, but sometimes is not good team member.	Usually listen to, shares with, and supports the efforts of others. Usually, respectful and listening actively	Almost always listens to, shares with and supports the efforts of others. Tries to keep people working well together.

Lab Report Rubric: *must be submitted in next lab***.**

Marks	1	2	3	4
Lab Report	The lab report does not follow the guidelines for formatting.	Presents some sections of the lab in the correct order. Three or more sections are not in the correct order; missing heading or title;	Presents most sections of the lab in the correct order, one or two sections may not be in the correct order; heading or title missing or not complete;	Presents all the sections of the lab in the correct order with correct formatting: includes correct heading, section headings and title of lab;

Lab Instructor: Engr. Zia Ur Rehman

Dr. A. Q. Khan Institute of Computer Sciences & Information Technology, (KICSIT)
Department of Computer Engineering

Lab Instructor: Engr. Zia Ur Rehman