

INTRODUCTION TO BIG DATA

ECAP456

Dr. Rajni Bhalla
Associate Professor

Learning Outcomes



After this lecture, you will be able to

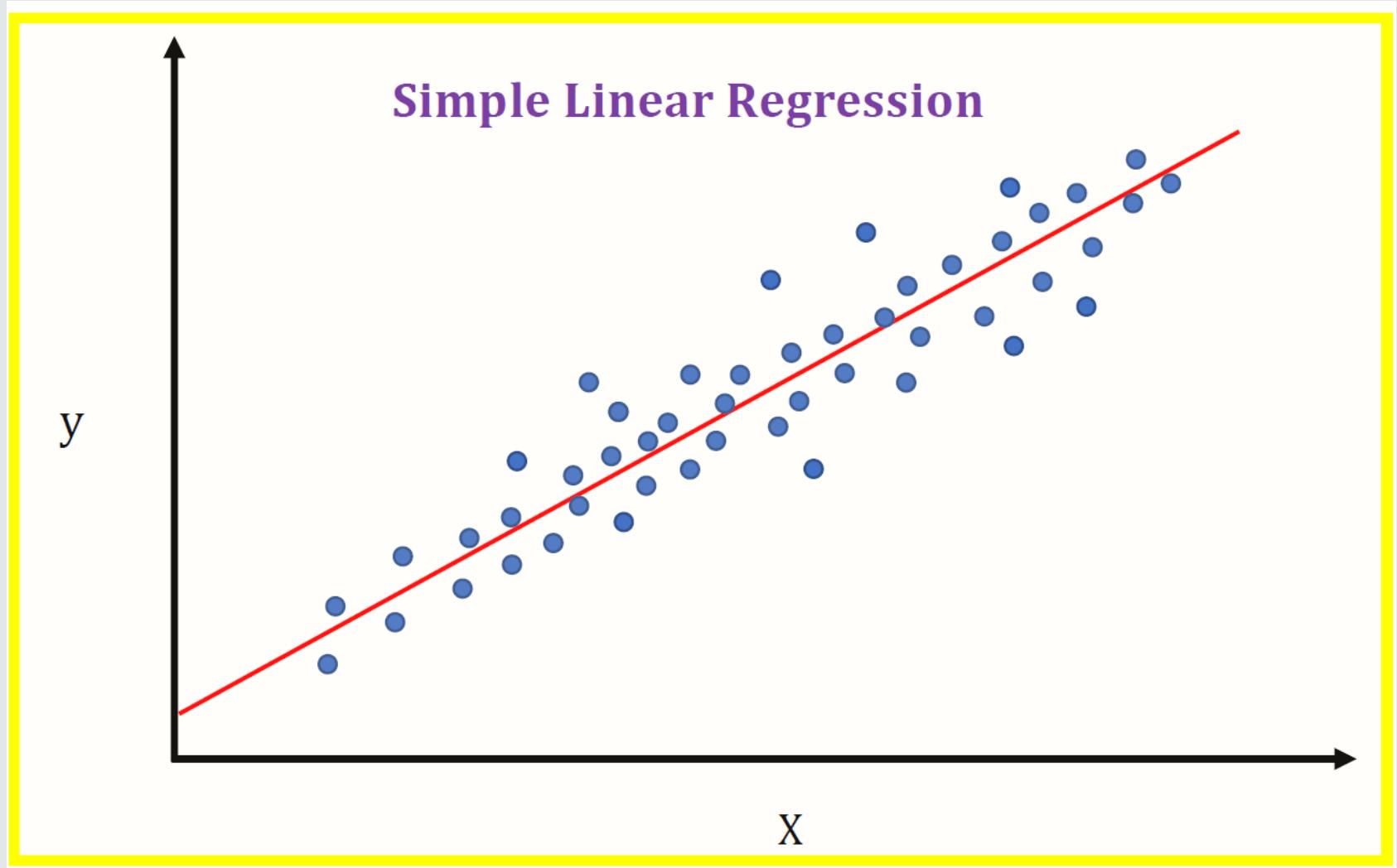
- Learn simple linear regression and multiple linear regression

Simple Linear Regression

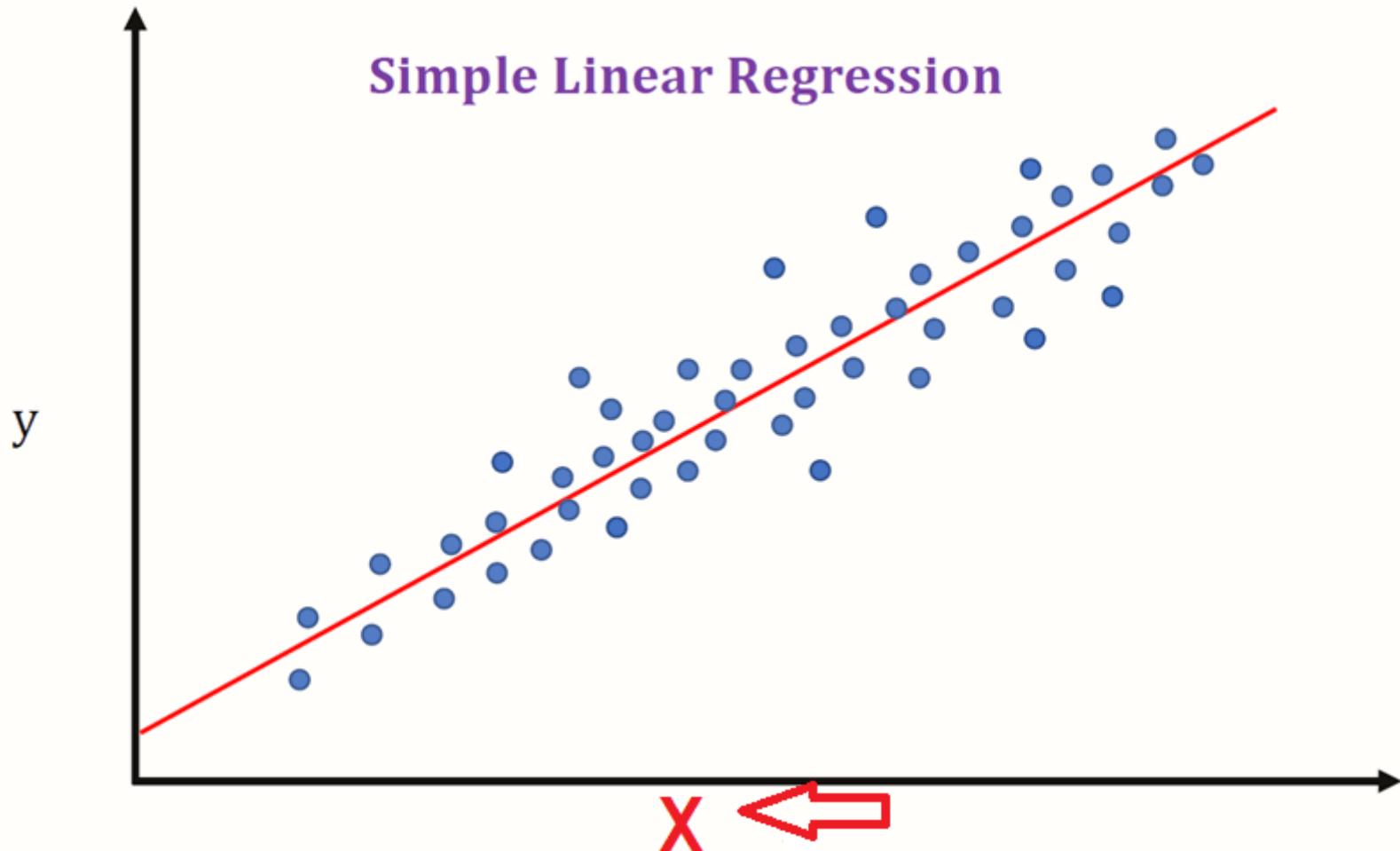
Simple Linear Regression

Simple Linear Regression

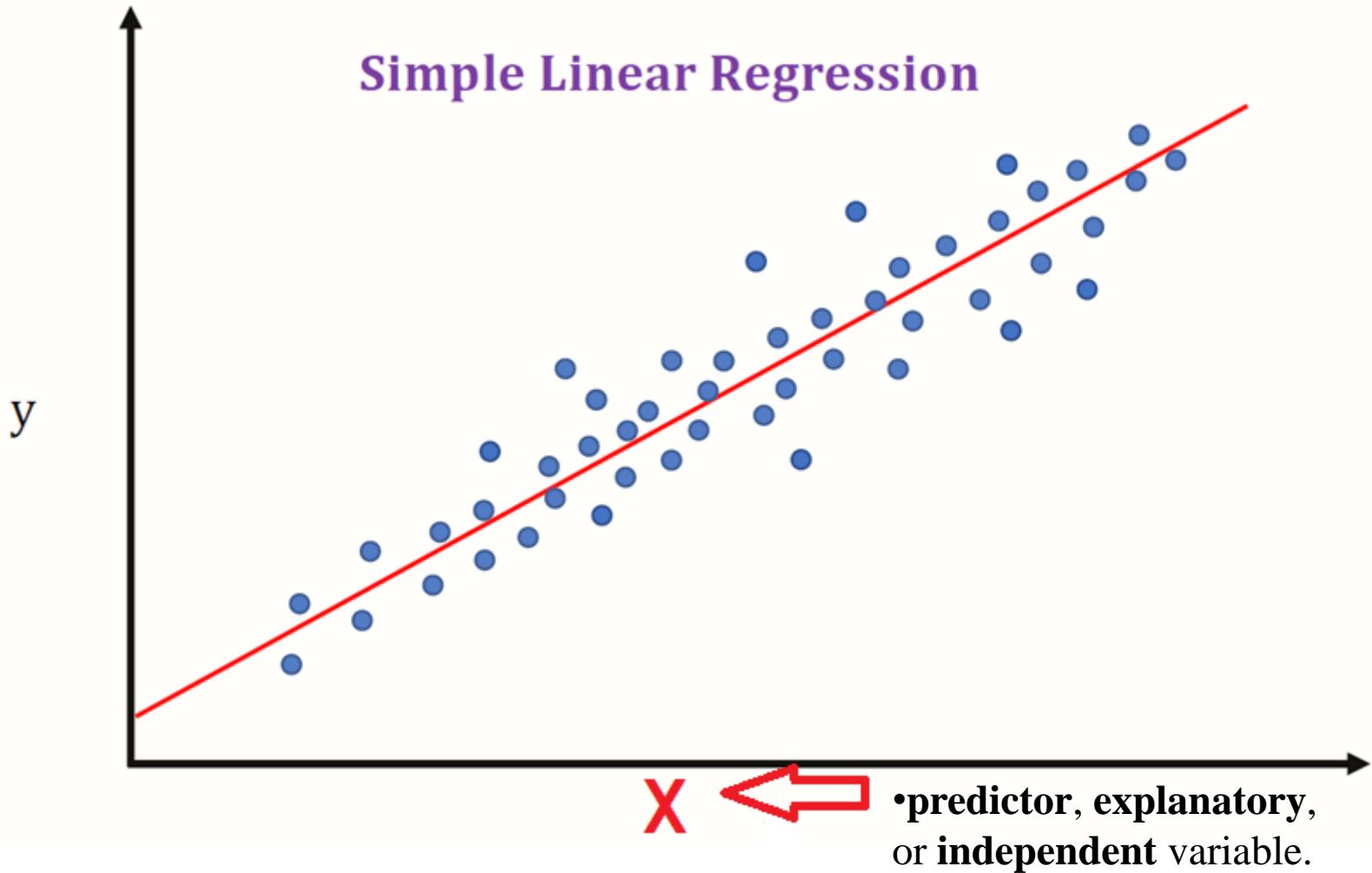
Introduction



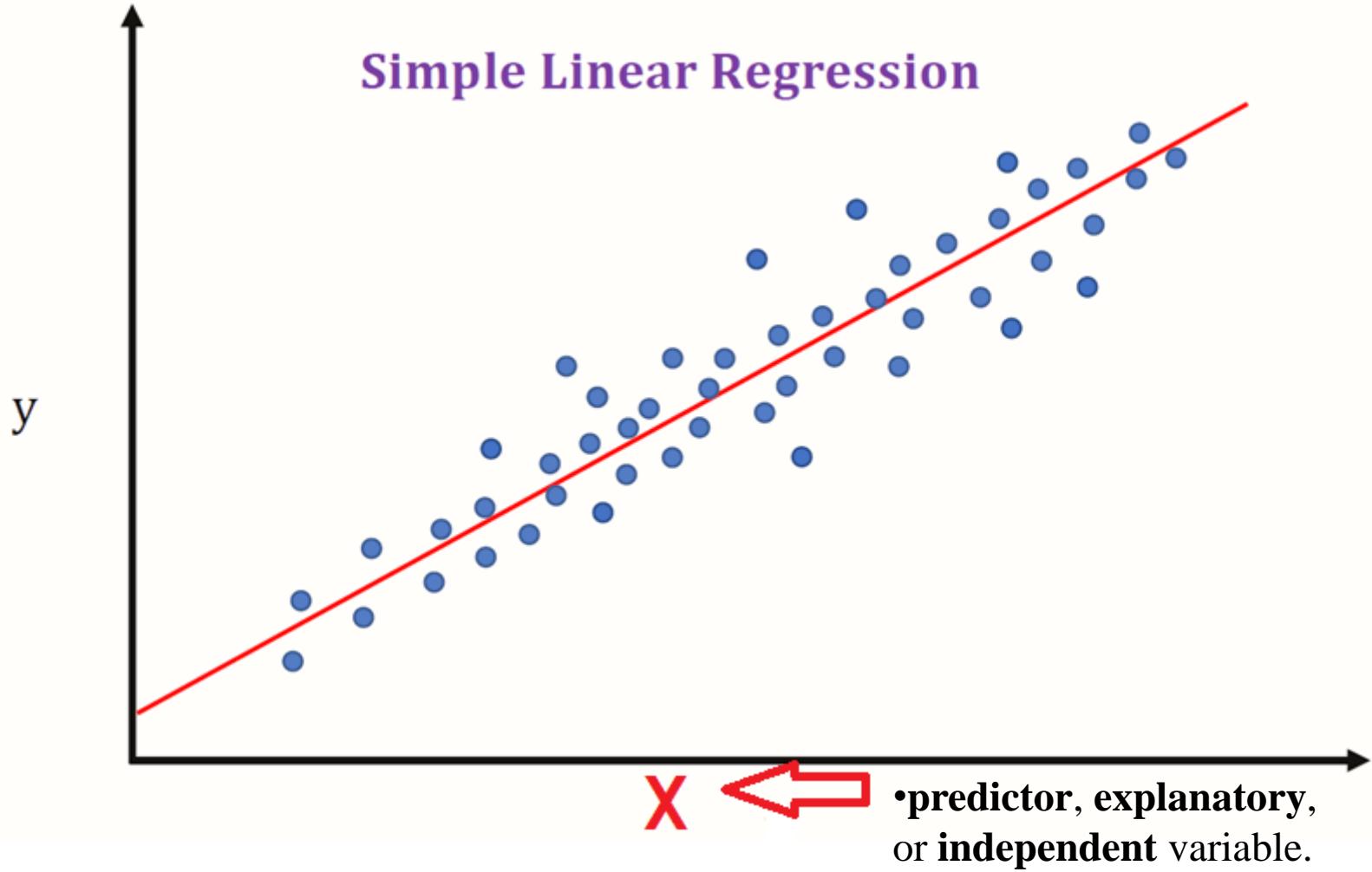
Introduction



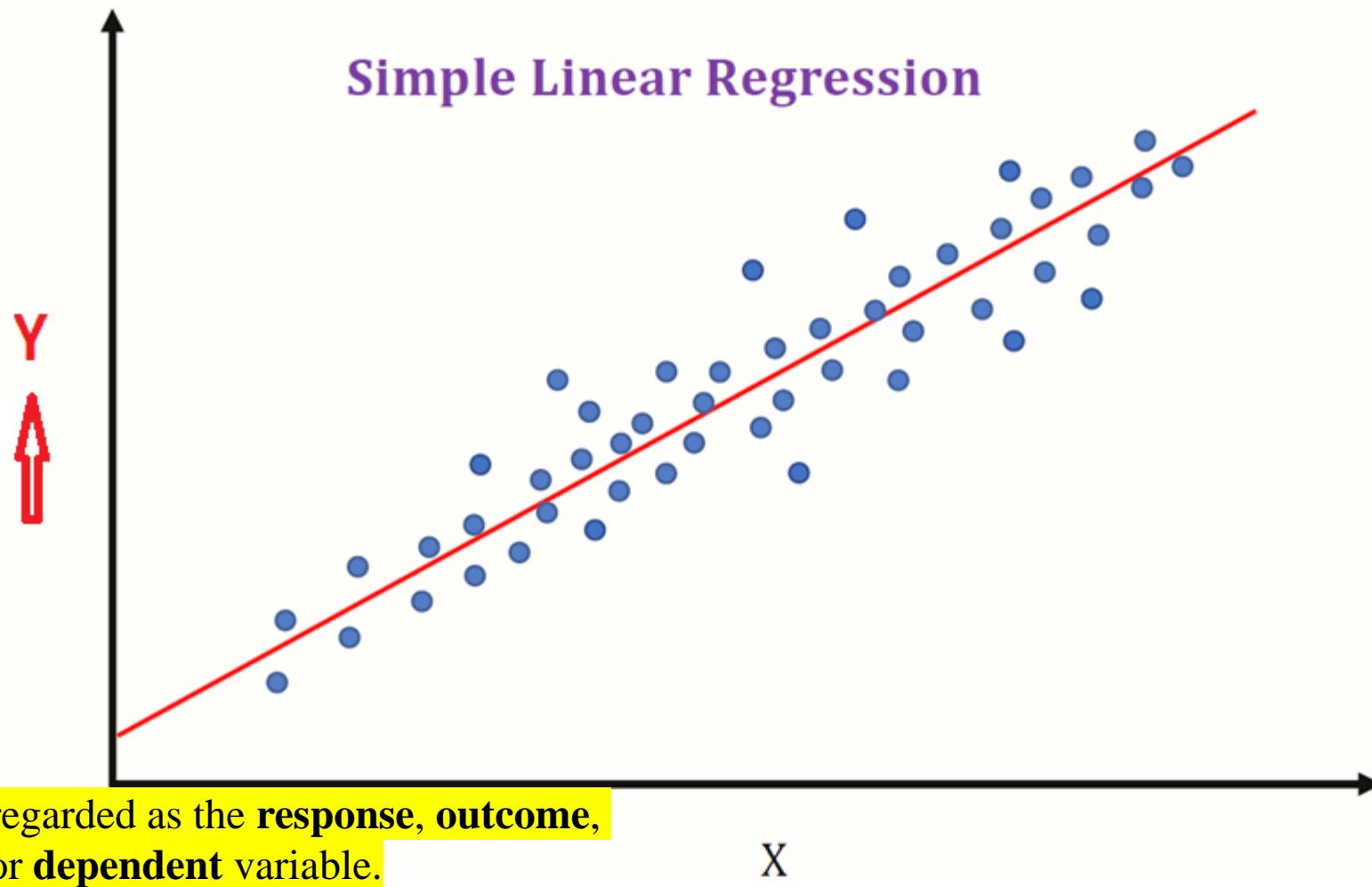
Introduction



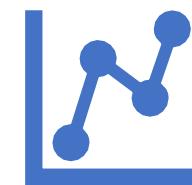
Introduction



Introduction



What is the objective of Linear regression ?



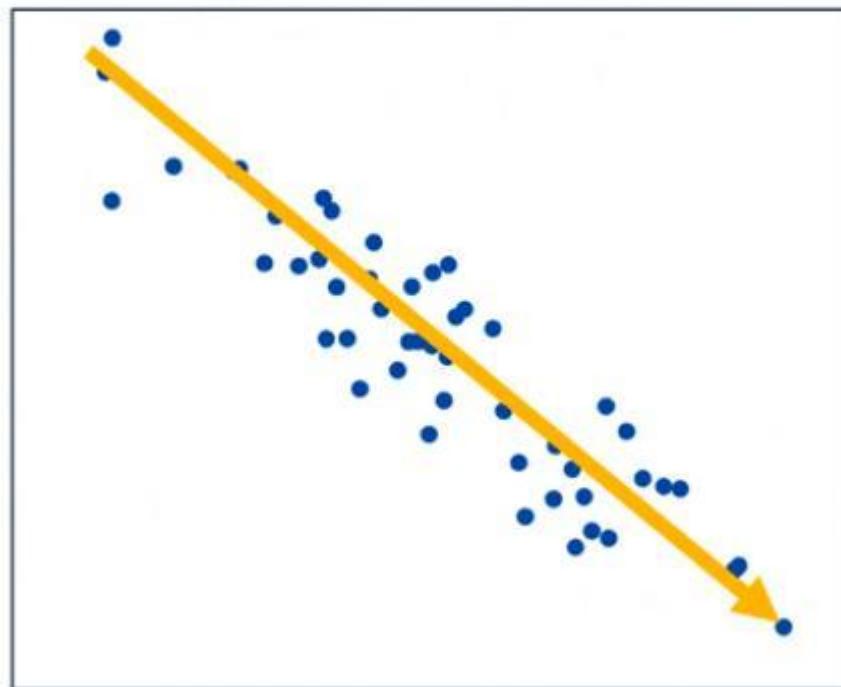
When to use regression ?

positive relationship



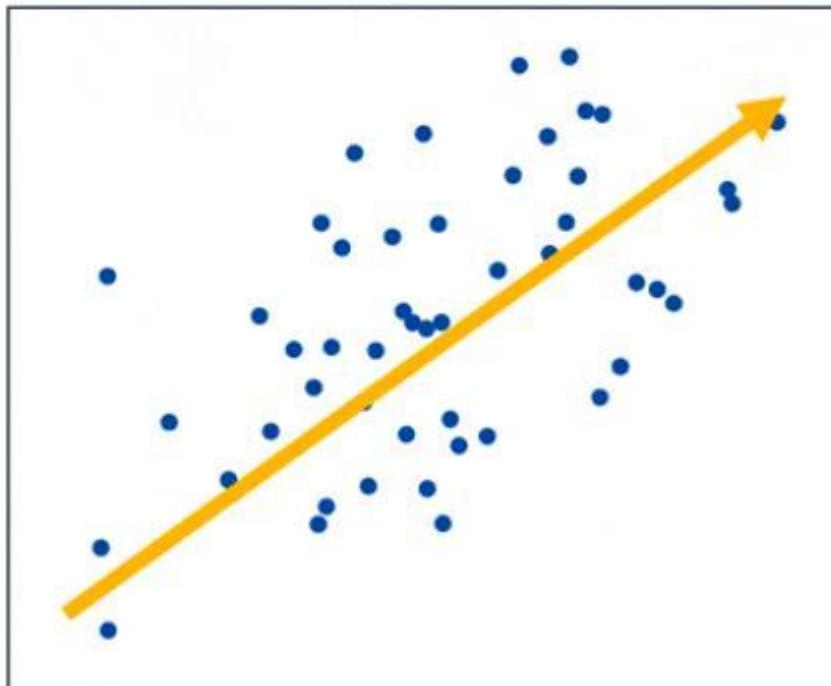
When to use regression ?

negative relationship



When to use regression ?

linear relationship



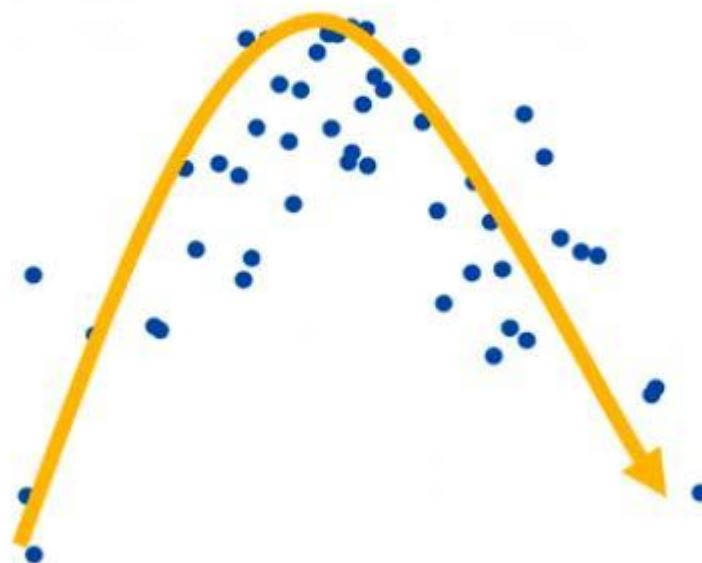
When to use regression ?

Zero



When to use regression ?

curvilinear relationship



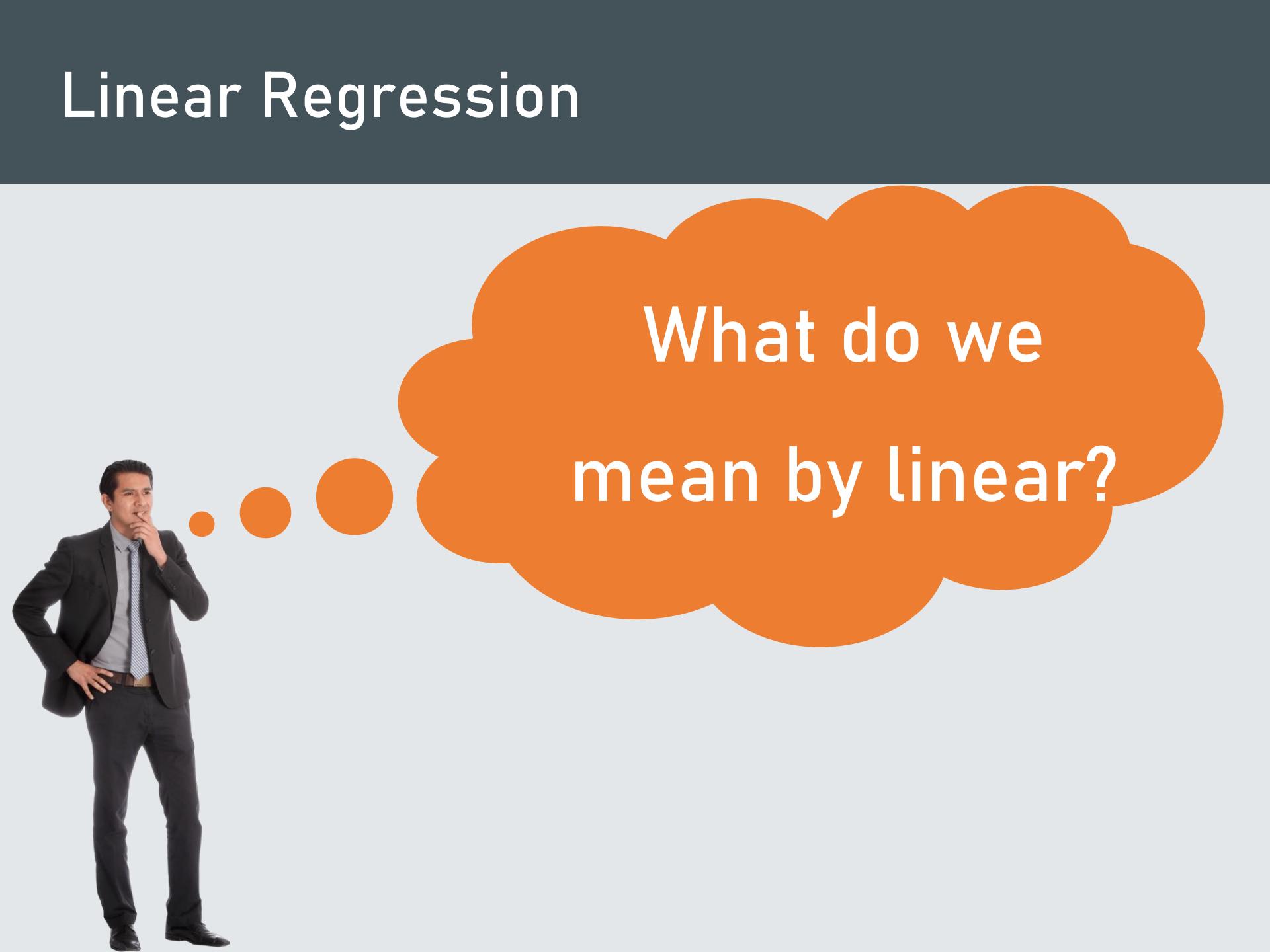
Linear Regression

In the context of Statistical learning, there are two types of data

Independent variables

Dependent variables:

Linear Regression

A man in a dark suit and tie stands on the left side of the frame, looking thoughtful with his hand to his chin. A large orange thought bubble originates from his head and extends towards the right. Inside the bubble, the text "What do we mean by linear?" is written in white, sans-serif font.

What do we
mean by linear?

Linear Regression

- Linear Equation

$$y = mx + c$$

y is the dependent variable.

x is the independent variable

m is the slope

c is the intercept

Linear Regression

- models are not perfect.
- approximate the relationship between dependent and independent variables
- Approximation leads to errors.
- irreducible error

Linear Regression

The same equation of a line can be re-written as:

$$Y = \beta_0 + \beta_1 X + \epsilon$$

Linear Regression

Example

make	fuelType	nDoors	engineSize	price
alfa-romero	gas	two	130	13495
alfa-romero	gas	two	130	16500
alfa-romero	gas	two	152	16500
audi	gas	four	109	13950
audi	gas	four	136	17450
audi	gas	two	136	15250
audi	gas	four	136	17710
audi	gas	four	136	18920
audi	gas	four	131	23875

Linear Regression

The following questions are addressed in the first set of analyses:

- Is price of car price related with engine size?
- How strong is the relationship?
- Is the relationship linear?
- Can we predict/estimate car price based on engine size?

Linear Regression

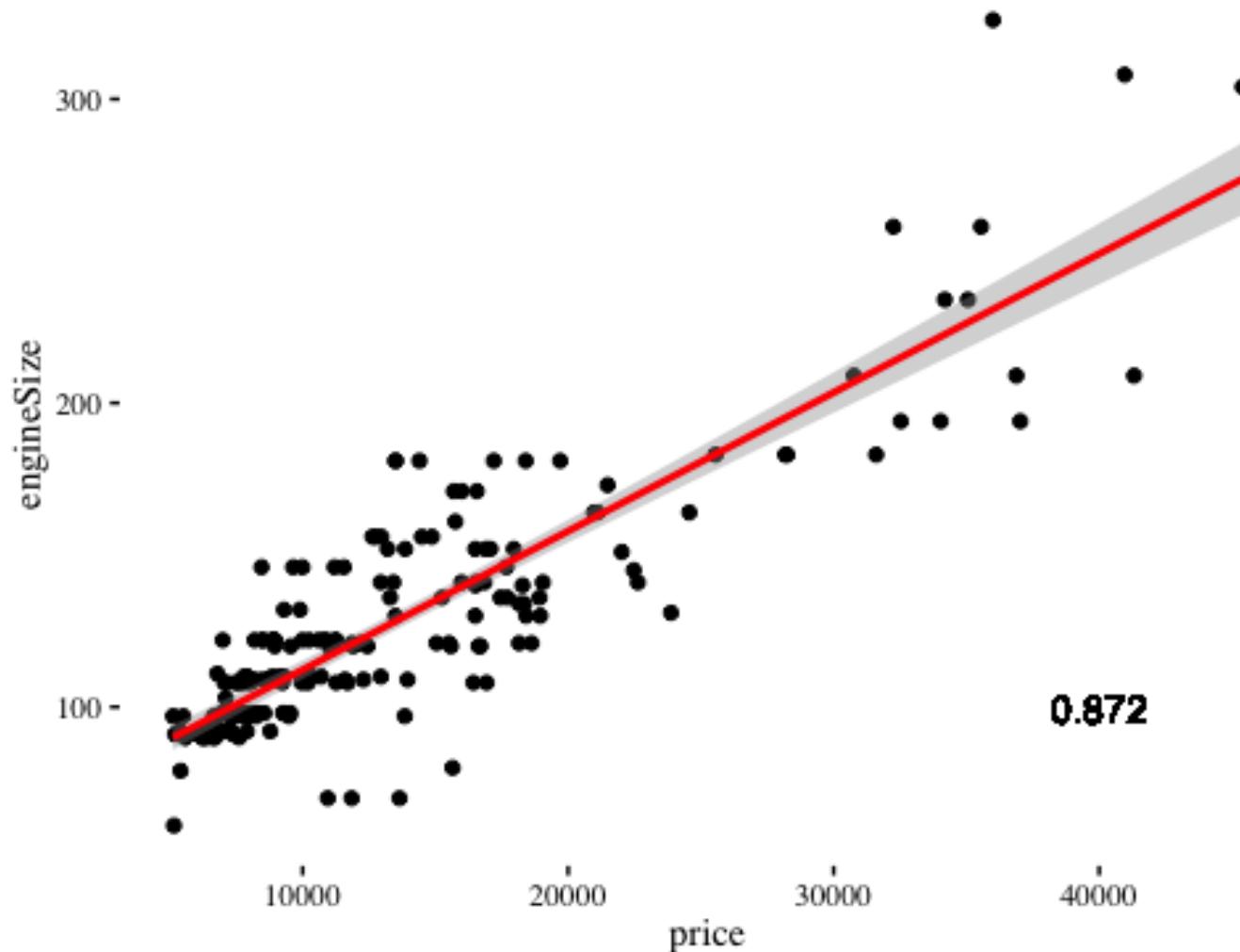
Fernando does a correlation study.

- Determining how closely two variables are linked.
- Its value ranges from 0 to 1.
- When the correlation coefficient is large (> 0.7), _____
- The presence of a big -ve number _____

Linear Regression

- He plots the relationship between price and engine size.
- He splits the data into training and test set.
- 75% of data is used for training. Remaining is used for the test.

Linear Regression



Linear Regression

Following are the answers to the questions:

- Is price of car price related with engine size?
- Yes, there is a relationship.

Linear Regression

Following are the answers to the questions:

- Is the relationship linear?
- A straight line can fit => A decent prediction of price can be made using engine size.

Linear Regression

Following are the answers to the questions:

- Can we predict/estimate the car price based on engine size?
- Yes, car price can be estimated based on engine size.

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
 - The value of the dependent variable at a certain value of the independent variables

Assumptions of multiple linear regression

Homogeneity of variance (homoscedasticity):

- Independence of observations
- Normality
- Linearity

How to perform a multiple linear regression

How to perform a multiple linear regression



That's all for now...