Simple Linear Regression

DATA SCIENCE | CCDATSCL

Linear regression or simply regression is a statistical model used to predict the relationship between **independent** and **dependent** variables.

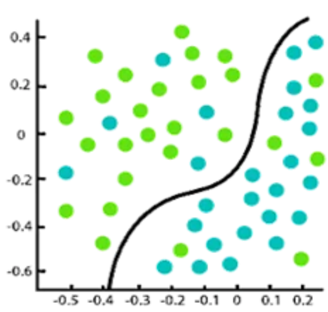
Independent Variable	Dependent Variable
A variable whose value does not change by the effect of other variables and is used to manipulate the **dependent variable**. Often denoted as **X**	A variable whose value changes when there is a manipulation in the values of the **independent variables**. Often denoted as **Y**

In Linear Regression, we examine **two factors.**

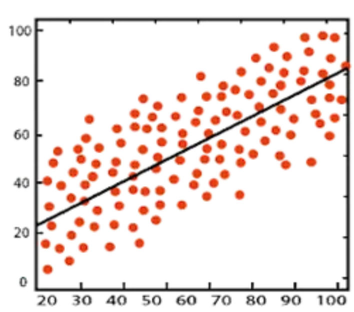
1. Which variables are significant predictors of the outcome variable?
2. How significant is the **regression line** in terms of making predictions with the highest possible accuracy?

Classification vs Regression

Classification	Regression
The output variable must be a **discrete value in the form of a class label**.	The output variable must be either **continuous in nature or a real value in the form of an integer quantity**.
Classification algorithms solve classification problems like identifying **spam e-mails, and spotting cancer cells** .	It is used to solve problems such as predicting **house prices and weather predictions**.
Classification tries to find the decision boundary, which **divides the dataset into different classes.**	It attempt to find the **best fit line**, which predicts the output more accurately.

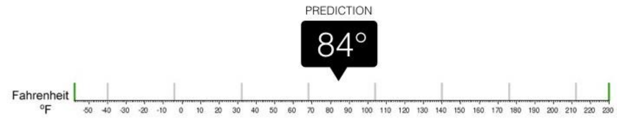


Classification

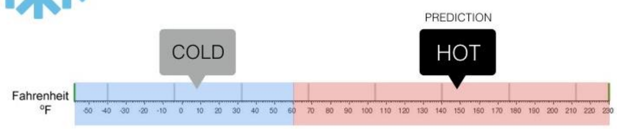


Regression

Regression
What is the temperature going to be tomorrow?

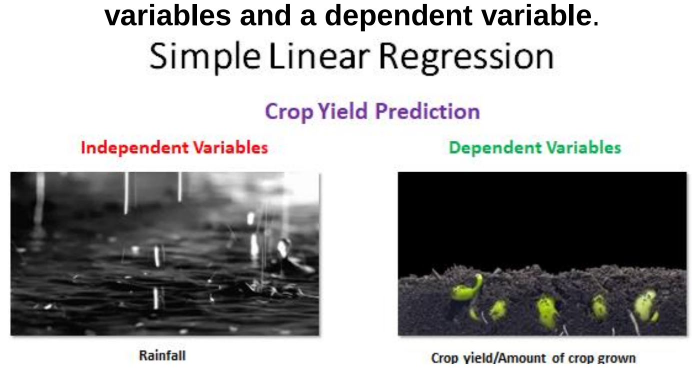


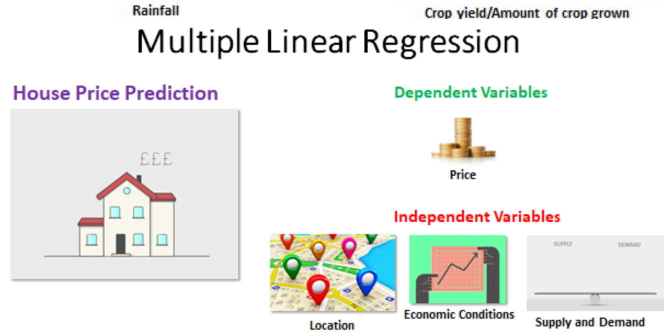
Classification
Will it be Cold or Hot tomorrow?

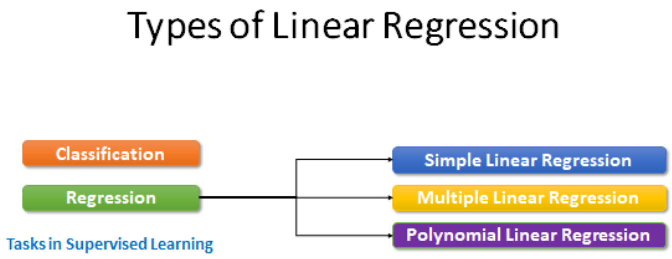


Types of Linear Regression

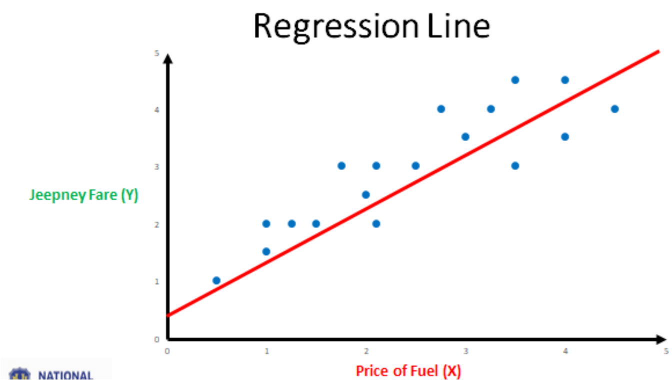
- **Simple Linear Regression**
  - This type involves estimating the relationship between **two quantitative variables**, such as the value of a dependent variable at a particular value of the independent variable.
- **Multiple Linear Regression**
  - In this type, you can determine the relationship between **several independent variables and a dependent variable**.





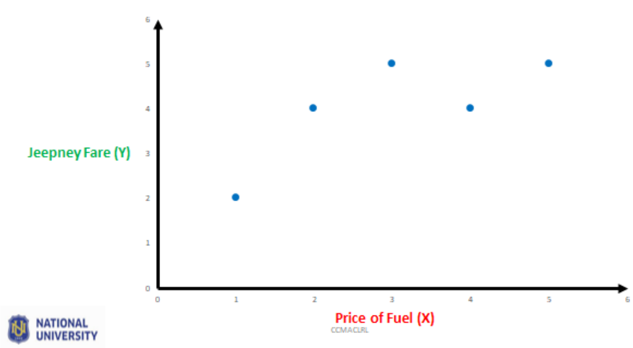


Regression Line



Least Squares Method

Price of Fuel (X)	Jeepney Fare (Y)
1	2
2	4
3	5
4	4
5	5



Price of Fuel (X)	Jeepney Fare (Y)	(X * Y)	X^2
1	2	2	1
2	4	8	4
3	5	15	9
4	4	16	16
5	5	25	25

$\Sigma x = 15$ $\Sigma y = 20$ $\Sigma xy = 66$ $\Sigma x^2 = 55$

Equation of the Line

The equation of the Line is given by this formula:

$\mathrm{Y} = m(x) + b$

where:

Y is the value of the **dependent variable**

X is the value of the **independent variable**

m is the **slope** of the line

b is the **y-intercept**

Calculating the **slope** is given by this formula:

$$m = \frac{n\,(\Sigma xy) - \Sigma x\,\Sigma y}{n\,(\Sigma x^2) - (\Sigma x)^2}$$

Calculating the **intercept** is given by this formula:

$$b = \frac{\Sigma y - m(\Sigma x)}{n}$$

Calculate the Slope

Σx	Σy	Σxy	Σx^2
15	20	66	55

$$m = \frac{n\,\Sigma xy \;-\; \Sigma x\,\Sigma y}{n\,\Sigma x2 \;-\; (\Sigma x)2}$$

$$m = \frac{5\,(66) \;-\; (15)\,(20)}{5\,(55) \;-\; (15)2}$$

$$m = 0.6$$

Calculate the Intercept

Σx	Σy	Σxy	Σx^2
15	20	66	55

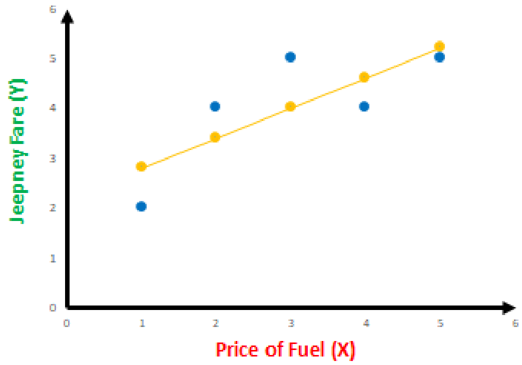
$$b = \frac{\Sigma y - m\,(\Sigma x)}{n}$$

$$b = \frac{20 - 0.6\,(15)}{5}$$

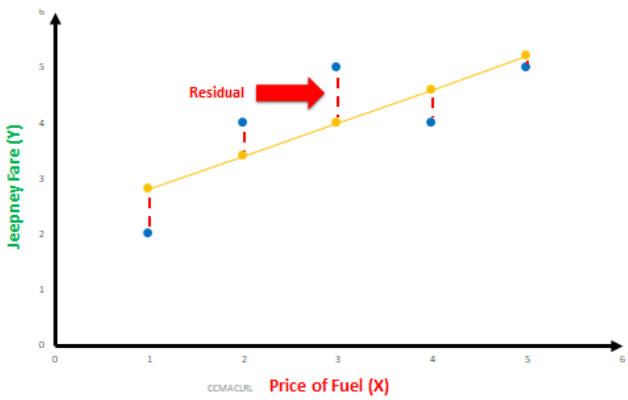
$$b = 2.2$$

Price of Fuel (X)	Jeepney Fare (Y)	Predicted Jeepney Fare ($Y_{predict}$)
1	2	2.8
2	4	3.4
3	5	4
4	4	4.6
5	5	5.2

$\mathbf{Y} = \boldsymbol{m}(\boldsymbol{x}) + \boldsymbol{b}$
$\mathbf{Y} = \mathbf{0.6(2) + (2.2)}$
$\mathbf{Y} = \mathbf{3.4}$
$\mathbf{Y} = \mathbf{0.6(5) + (2.2)}$
$\mathbf{Y} = \mathbf{5.2}$



The **blue points** represent **the actual Y values** and **the orange points** represent the **predicted Y values**



The **distance** between the **actual values** and the **predicted values** are known as **residuals or errors**

Loss Function

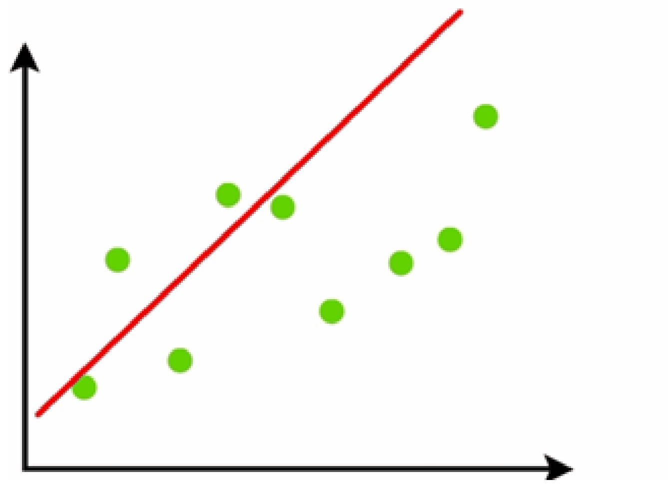
Price of Fuel (X)	Jeepney Fare (Y)	Predicted Jeepney Fare ($Y_{predict}$)	Y - $Y_{predict}$	$(Y - Y_{predict})^2$
1	2	2.8	-0.8	0.64
2	4	3.4	0.6	0.36
3	5	4	1	1
4	4	4.6	-0.6	0.36
5	5	5.2	-0.2	0.04

$$\boldsymbol{Sum\ of\ Squared\ Errors\ (SSE)} = \sum_{i=1}^{n}(y_i - y_{predict})^2$$

The sum of squared errors for this regression line is **2.4**. This tells you how **good a line is fitted to the data**. **The best fit line will have the least amount of this value.**
SSE = 2.4

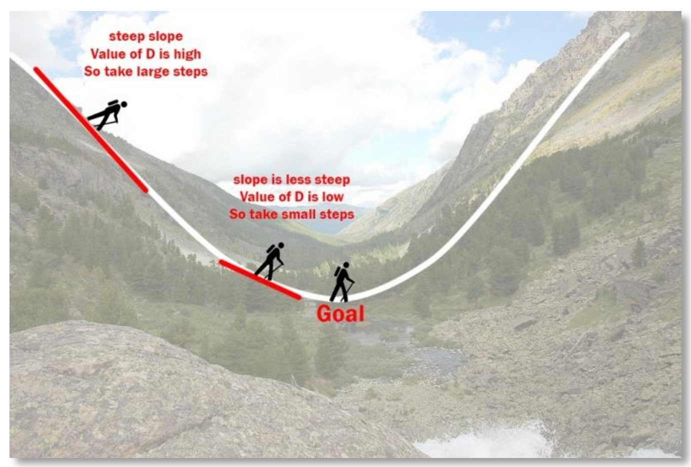
Finding the Best Fit Regression Line

We keep moving this line through the **data points** to make sure the best fit line **has the least square distance** between the **data points** and the **regression line**



Gradient Descent

Gradient descent is an iterative optimization algorithm to find the **minimum of a function**. Here that function is our **Loss Function**.



When going down a valley,
m (slope) is the current position of the person
D (partial derivative) is the steepness of the slope
L (Learning Rate) is the speed at which the person moves

L x D be the size of the steps the person will take

When the slope is more steep he takes **longer steps** and when it is less steep, he takes **smaller steps**.
Finally he arrives at the bottom of the valley which corresponds to our **loss = 0**.

$$D_m = \frac{1}{n} \sum_{i=0}^{n} 2\left(yi - (mxi + c)\right)(-xi)$$

$$D_m = \frac{-2}{n} \sum_{i=0}^{n} x_i \, (yi - \bar{y}i)$$

$$D_b = \frac{-2}{n} \sum_{i=0}^{n} (yi - \bar{y}i)$$

$$m = m - L * D_b$$
$$b = b - L * D_b$$