SIMPLE LINEAR REGRESSION

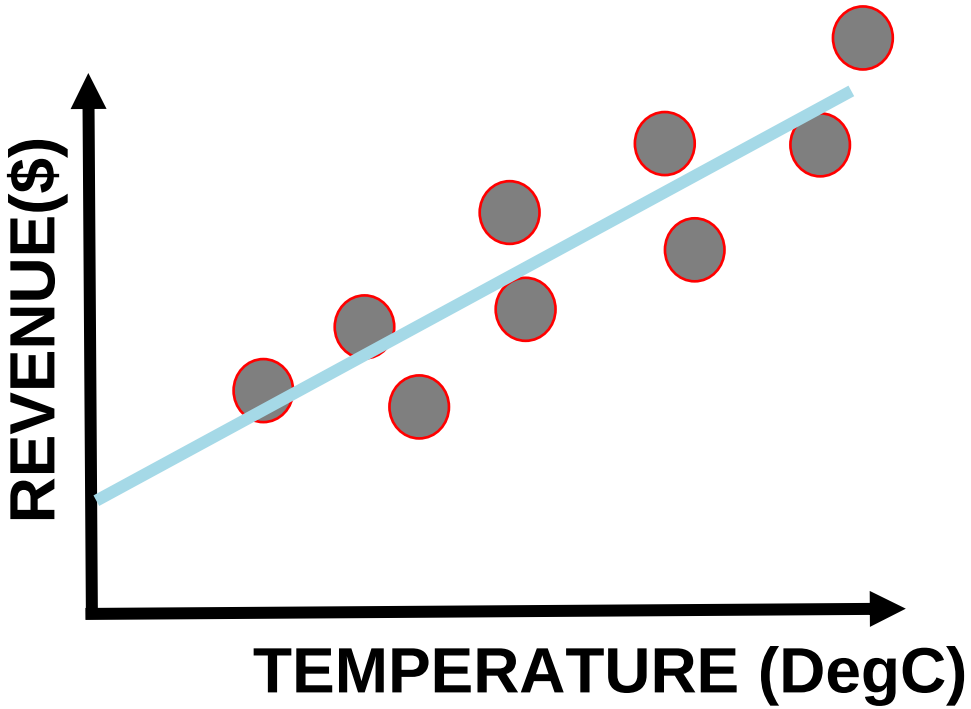
Prof. Venus R. Patel

PROJECT OVERVIEW

- You own an ice cream business and you would like to create a model that could predict the daily revenue in dollars based on the outside air temperature (degC).

- Dataset:
  - Input (X): Outside Air Temperature
  - Output (Y): Overall daily revenue generated in dollars

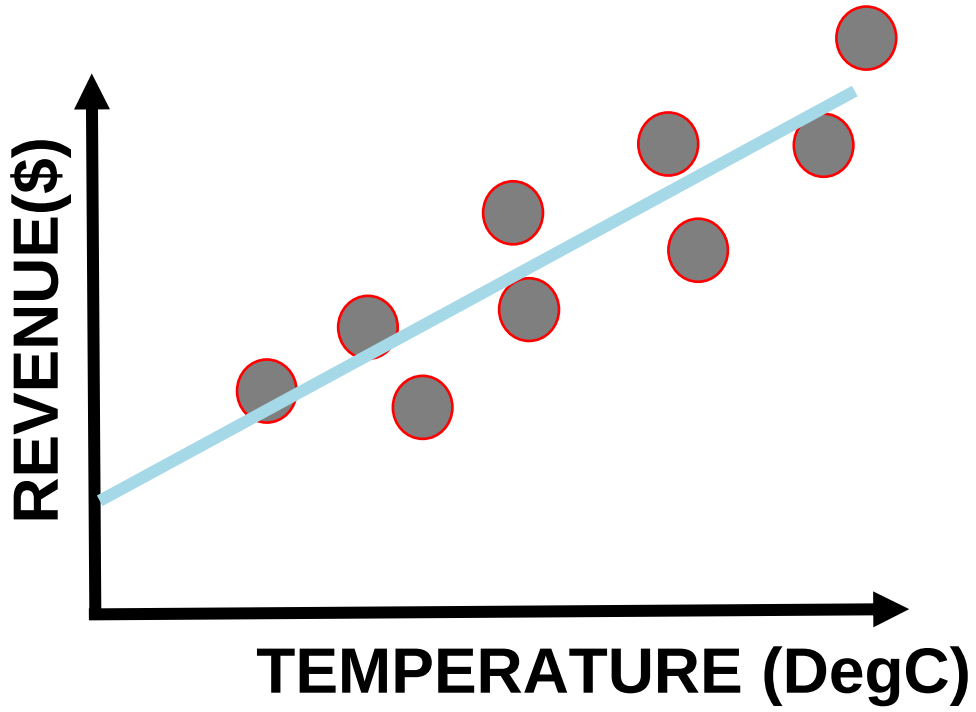
	Temperature	Revenue
0	24.566884	534.799028
1	26.005191	625.190122
2	27.790554	660.632289
3	20.595335	487.706960
4	11.503498	316.240194
5	14.352514	367.940744
6	13.707780	308.894518
7	30.833985	696.716640
8	0.976870	55.390338
9	31.669465	737.800824
10	11.455253	325.968408
11	3.664670	71.160153



Source: https://www.goodfreephotos.com/vector-images/ice-cream-stand-vector-clipart.png.php

SIMPLE LINEAR REGRESSION

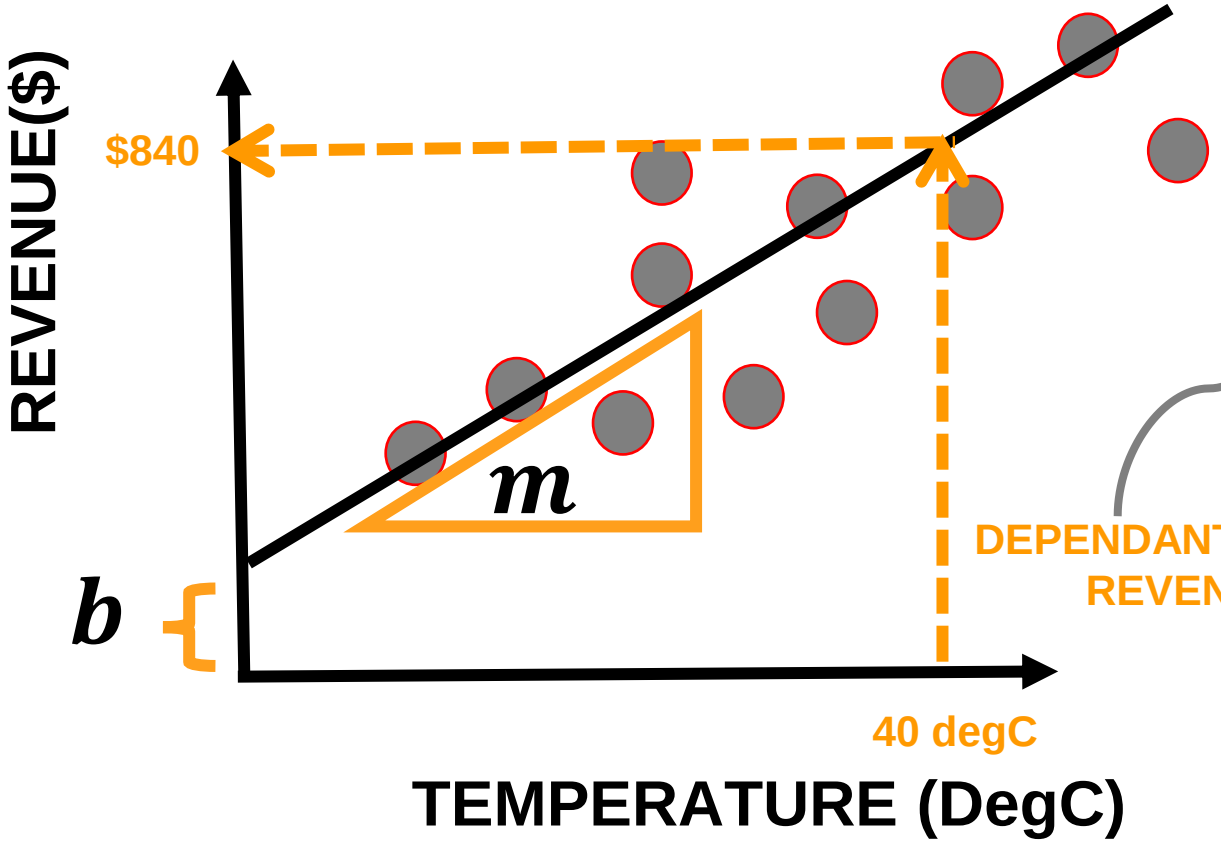
- In simple linear regression, we predict the value of one variable Y based on another variable X.
- X is called the independent variable and Y is called the dependent variable.
- Why simple? Because it examines the relationship between two variables only.
- Why linear? when the independent variable increases (or decreases), the dependent variable increases (or decreases) in a linear fashion.



	Temperature	Revenue
0	24.566884	534.799028
1	26.005191	625.190122
2	27.790554	660.632289
3	20.595335	487.706960
4	11.503498	316.240194
5	14.352514	367.940744
6	13.707780	308.894518
7	30.833985	696.716640
8	0.976870	55.390338
9	31.669465	737.800824
10	11.455253	325.968408
11	3.664670	71.160153

SIMPLE LINEAR REGRESSION: SOME MATH!

- Goal is to obtain a relationship (model) between outside air temperature and ice cream sales revenue. Simply you need to find "*m*" and "*b*".
- This "trained" model can be later used to predict any Revenue (dollars) based on the outside air Temperature.



MODEL! (GOAL)

$$y = b + m * x$$

DEPENDANT VARIABLE
REVENUE ($)

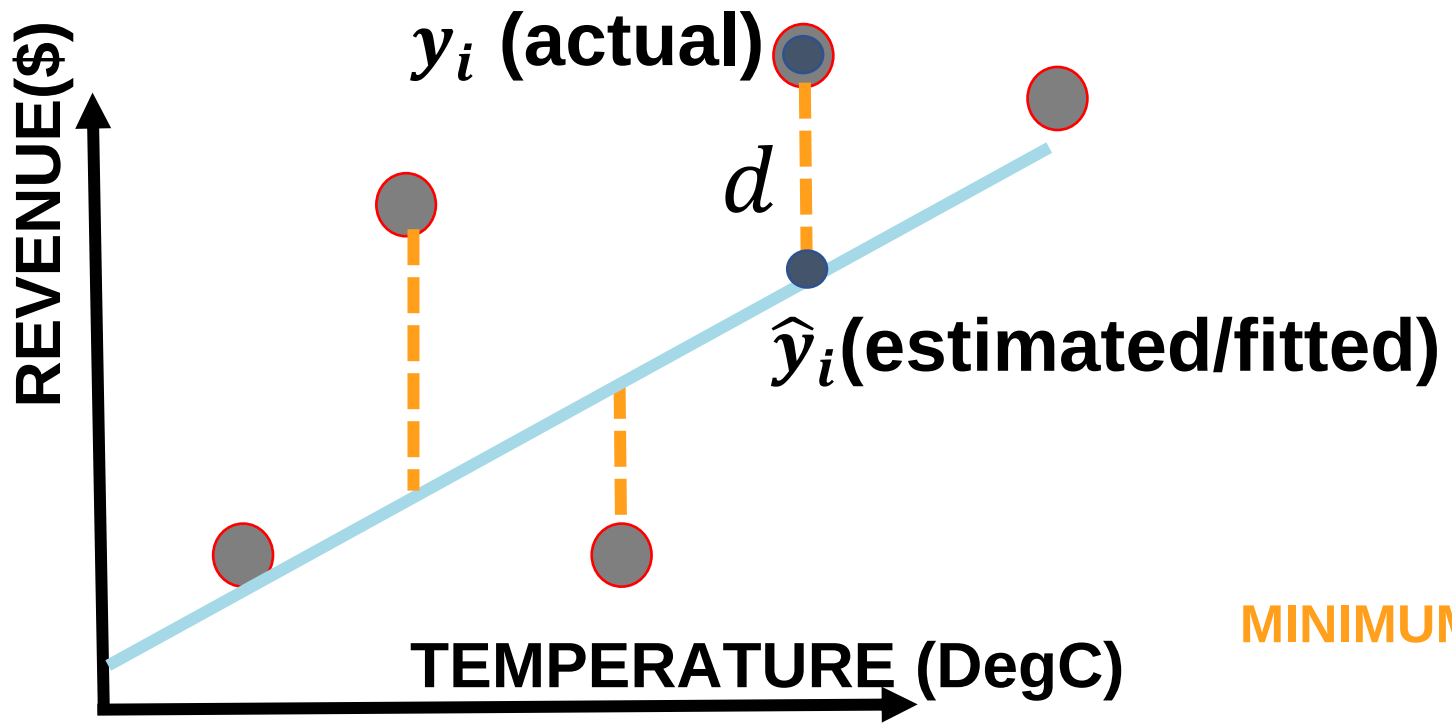
INDEPENDENT VARIABLE
TEMPERATURE (DegC)

Temperature (x) = 40 degC
Revenue (y) = $840

LEAST SUM OF SQUARES

HOW TO GET MODEL PARAMETERS?
LEAST SUM OF SQUARES

- Least squares fitting is a way to find the best-fit curve or line for a set of points.
- The sum of the squares of the offsets (residuals) is used to estimate the best-fit curve or line.
- Least squares method is used to obtain the coefficients m and b.



$$d = \hat{y}_i - y_i$$

$$min \sum (\hat{y}_i - y_i)^2$$

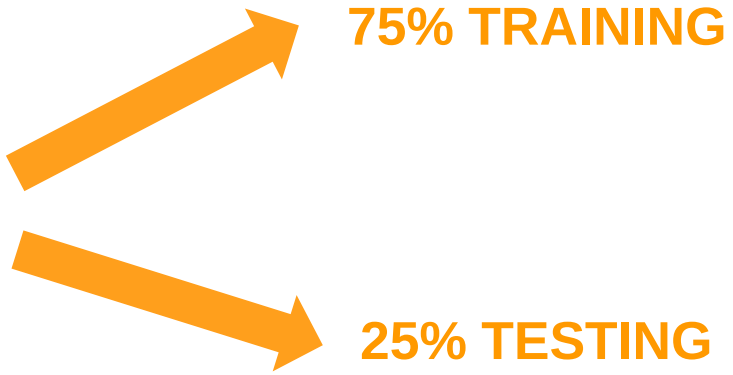
MINIMUM (LEAST) SUM OF SQUARES

TRAINING VS. TESTING DATASET

- Data set is divided into 75% for training and 25% for testing.
  - Training set: used for model training.
  - Testing set: used for testing trained model. Make sure that the testing dataset has never been seen by the trained model before.

TOTAL DATA

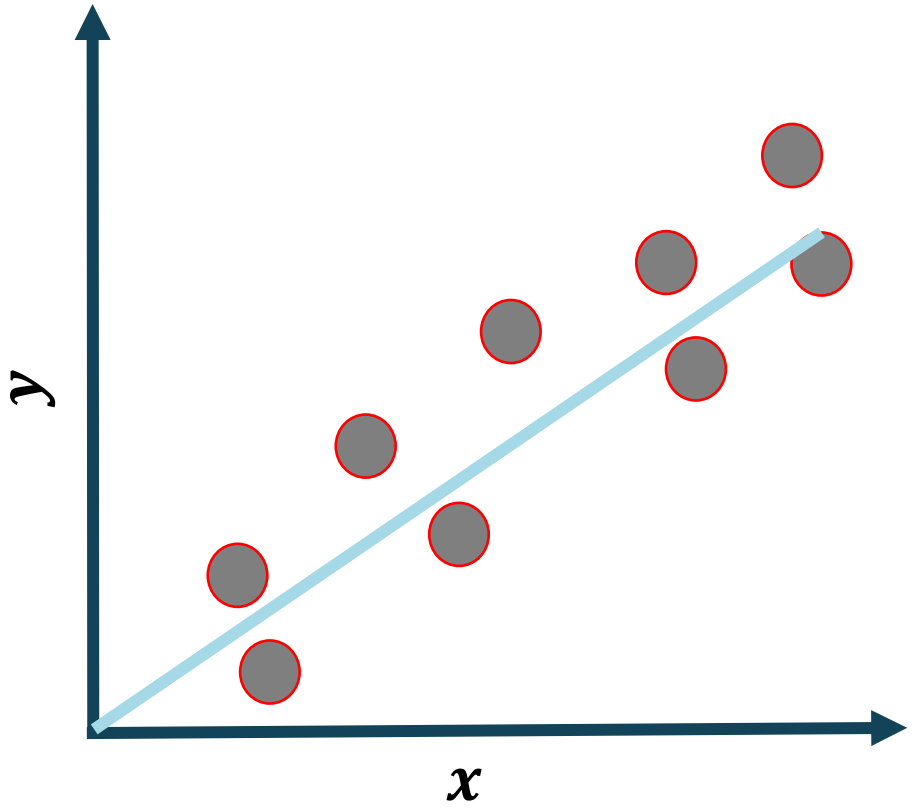
	Temperature	Revenue
0	24.566884	534.799028
1	26.005191	625.190122
2	27.790554	660.632289
3	20.595335	487.706960
4	11.503498	316.240194
5	14.352514	367.940744
6	13.707780	308.894518
7	30.833985	696.716640
8	0.976870	55.390338
9	31.669465	737.800824
10	11.455253	325.968408
11	3.664670	71.160153

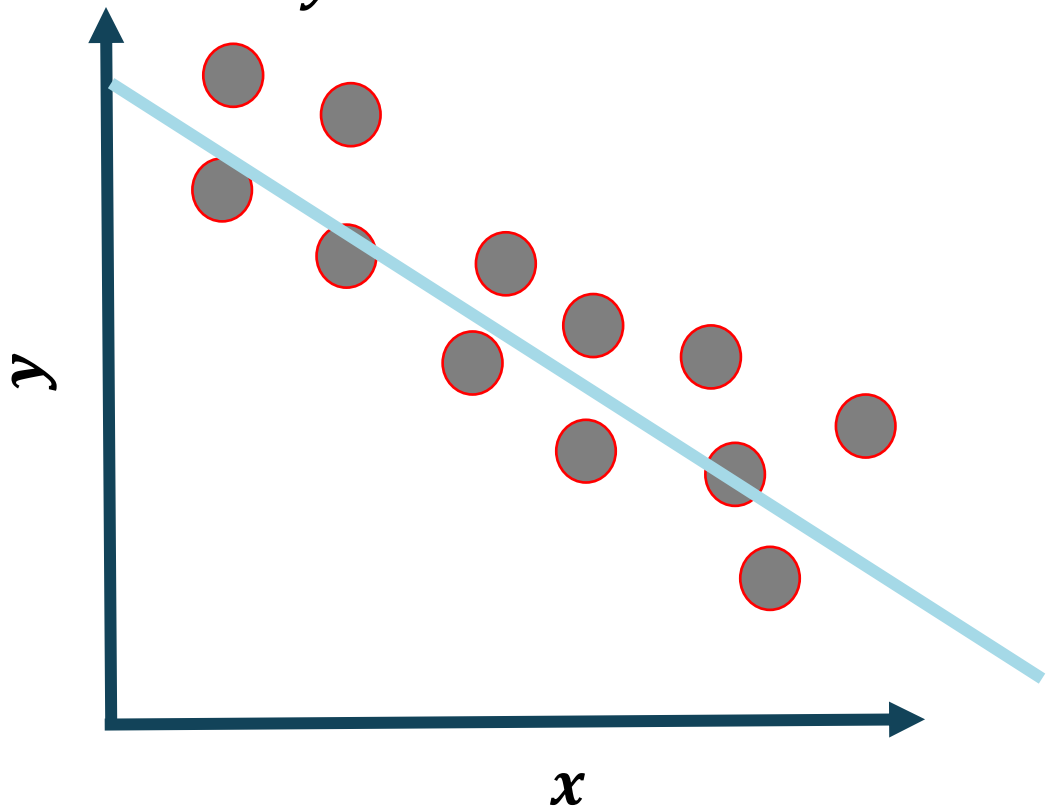


SIMPLE LINEAR REGRESSION: PRACTICE OPPORTUNITY

- Match the equations to the figures:

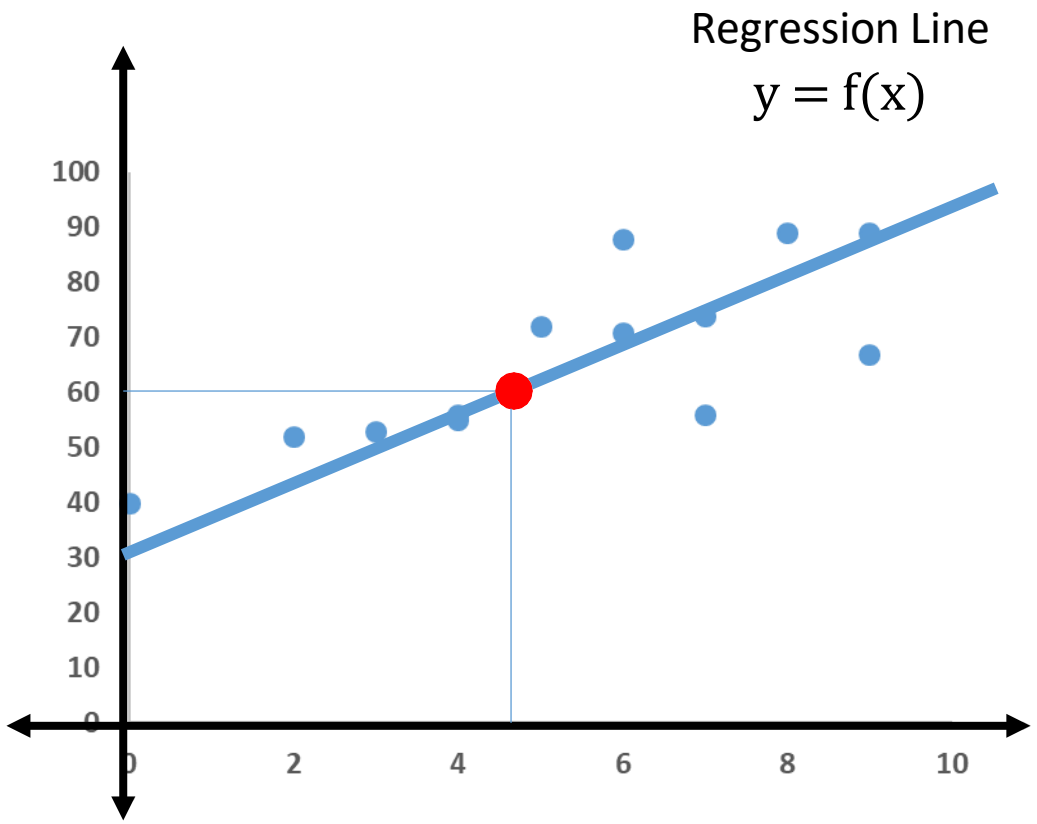
$y = 3 * x$ $y = 15 - 10 * x$





Regression Analysis

- Statistical process for estimating the relationships among variables
- The predictor is a continuous variable
- Relationship between a dependent variable and one or
  more independent variables (or 'predictors')
- Can also be used to infer causal relationships between
  dependent and independent variables.

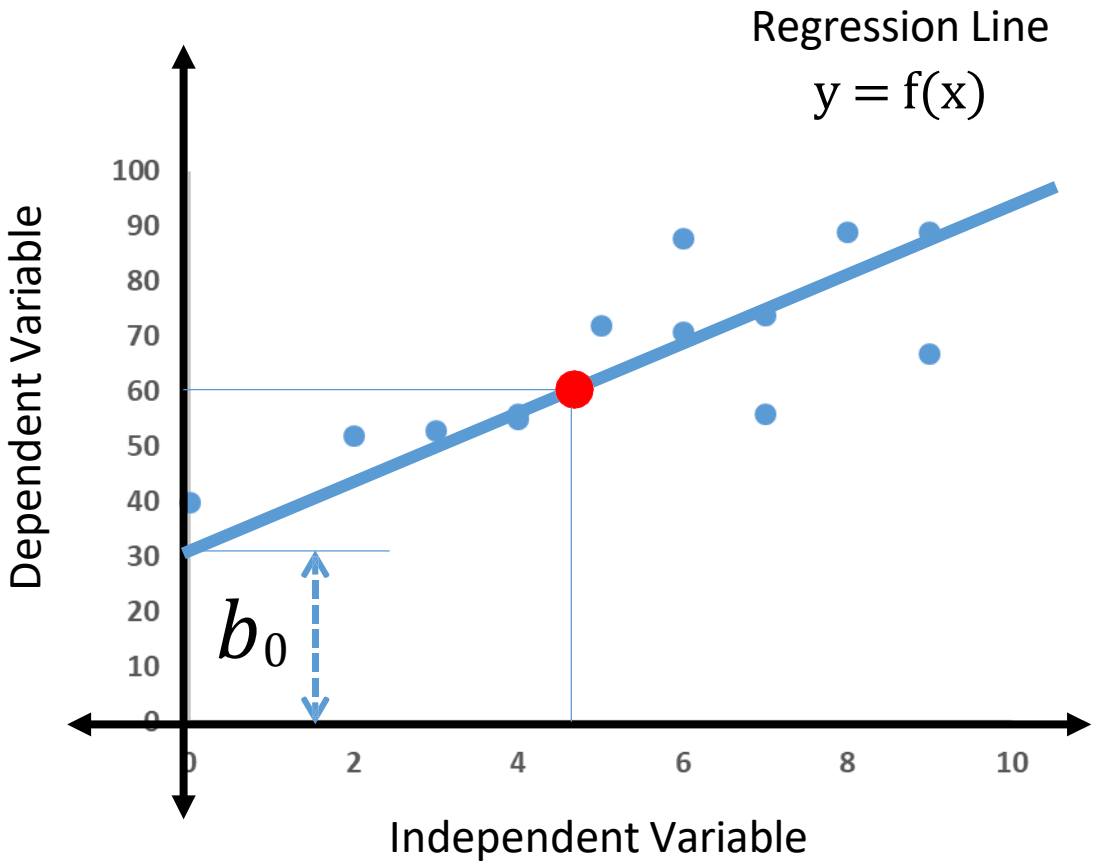


Simple Linear Regression

Simple Regression :

$$y = b_0 + b_1 x$$

Only one Dependent Only one Independent



Simple Linear Regression

Hrs Studied (X)	Marks (Y)
0	40
2	52
3	53
4	55
4	56
5	72
6	71
6	88
7	56
7	74
8	89
9	67
9	89
5.38	66.31
Mean	

X – Mean (A)	Y – Mean (B)	A^2	A*B
-5.38	-26.31	28.99	141.66
-3.38	-14.31	11.46	48.43
-2.38	-13.31	5.69	31.73
-1.38	-11.31	1.92	15.66
-1.38	-10.31	1.92	14.27
-0.38	5.69	0.15	-2.19
0.62	4.69	0.38	2.89
0.62	21.69	0.38	13.35
1.62	-10.31	2.61	-16.65
1.62	7.69	2.61	12.43
2.62	22.69	6.84	59.35
3.62	0.69	13.07	2.50
3.62	22.69	13.07	82.04
		89.08	405.46
		Sum	

$$y = b_0 + b_1 x$$

$$b_1 = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{\sum (X - \bar{X})^2}$$

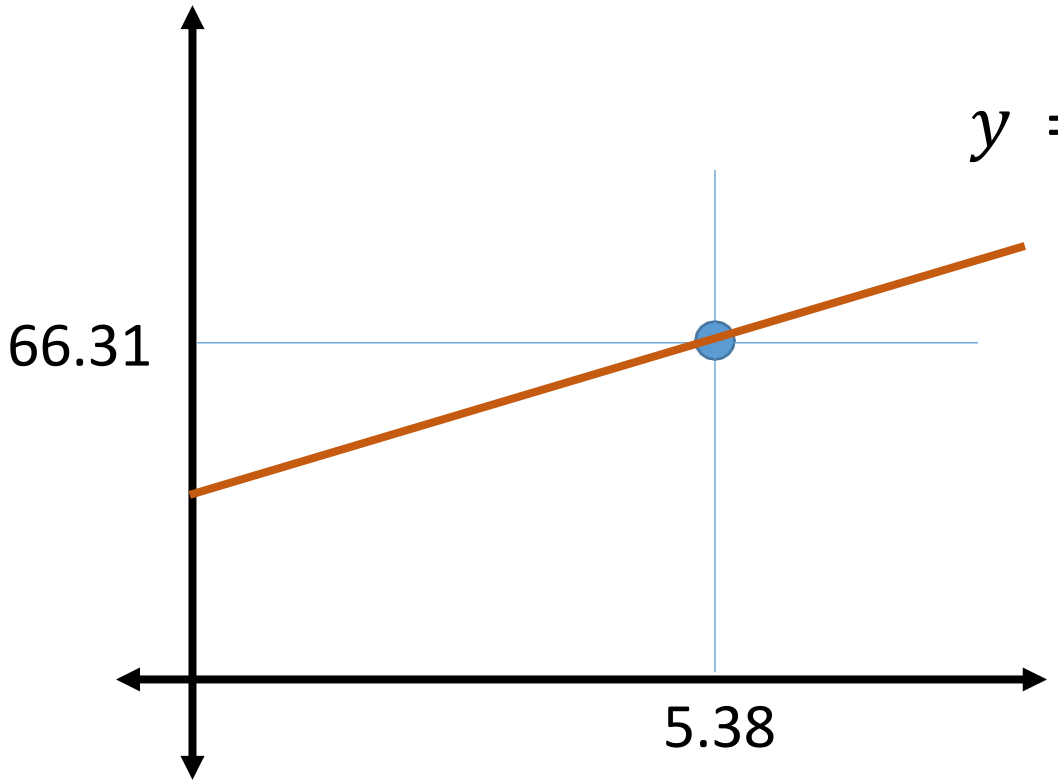
$$= 405.46 / 89.08$$

$$= 4.55$$

Simple Linear Regression

Hrs Studied (X)	Marks (Y)
0	40
2	52
3	53
4	55
4	56
5	72
6	71
6	88
7	56
7	74
8	89
9	67
9	89
5.38	66.31
Mean	

$y = b_0 + b_1 x$ $b_1 = 4.55$ $b_0 = 41.8$



$y = 41.8 + 4.55\,x$

SCIKIT-LEARN

SCIKIT-LEARN

- Scikit-learn is a free machine-learning library developed for python.
- Scikit-learn offers several algorithms for classification, regression, and clustering.
- Several famous models are included such as support vector machines, random forests, gradient boosting, and k-means.
- Scikit learn can be efficiently used in data preprocessing.