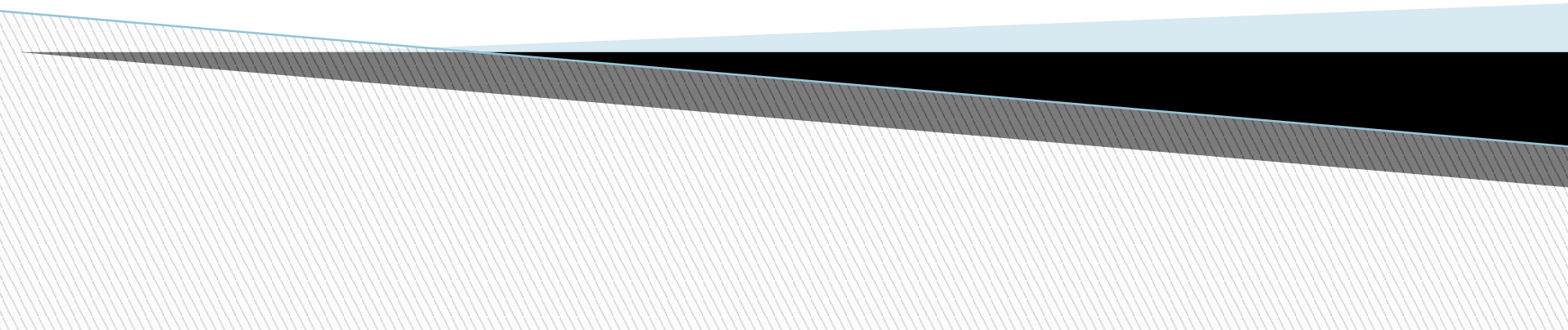


Univariate Linear Regression

Univariate linear regression focuses on determining **relationship between** one or more independent (explanatory variable) variable and one dependent variable.



Linear regression

- ❑ Linear regression makes predictions for continuous/real or numeric variables such as **sales, salary, age, product price**, etc.
- ❑ It is a supervised technique

Types of Linear Regression

- ❑ **Simple Linear Regression:**
If a single independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Simple Linear Regression.
- ❑ **Multiple Linear regression:**
If more than one independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Multiple Linear Regression.

YearsExperience	Salary
1.1	39343
1.3	46205
1.5	37731
2	43525
2.2	39891
2.9	56642
3	60150
3.2	54445
3.2	64445
3.7	57189
3.9	63218
4	55794
4	56957
4.1	57081
4.5	61111

Simple Linear Regression

Multiple Linear regression

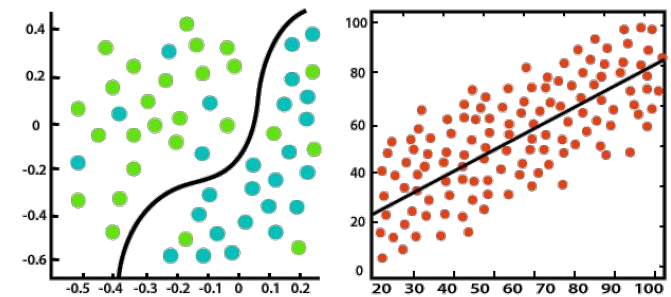
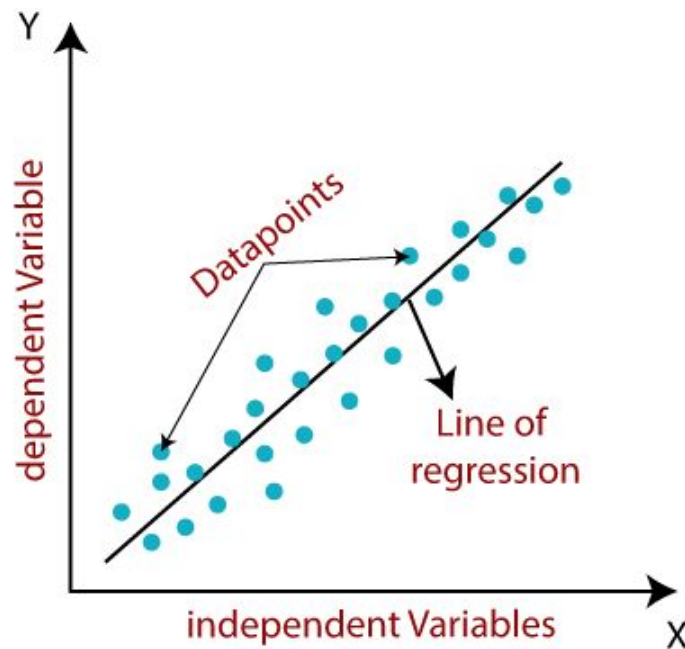
bedrooms	bathroom	sqft_living	sqft_lot	floors	waterfront	view	condition	sqft_above	sqft_base	yr_built	yr_renovated	street	city	statezip	country	price
3	1.5	1340	7912	1.5	0	0	3	1340	0	1955	2005	18810 Den	Shoreline	WA 98133	USA	313000
5	2.5	3650	9050	2	0	4	5	3370	280	1921	0	709 W Bl	Seattle	WA 98119	USA	2384000
3	2	1930	11947	1	0	0	4	1930	0	1966	0	26206-262	Kent	WA 98042	USA	342000
3	2.25	2000	8030	1	0	0	4	1000	1000	1963	0	857 170th	Bellevue	WA 98008	USA	420000
4	2.5	1940	10500	1	0	0	4	1140	800	1976	1992	9105 170th	Redmond	WA 98052	USA	550000
2	1	880	6380	1	0	0	3	880	0	1938	1994	522 NE 88	Seattle	WA 98115	USA	490000
2	2	1350	2560	1	0	0	3	1350	0	1976	0	2616 174th	Redmond	WA 98052	USA	335000
4	2.5	2710	35868	2	0	0	3	2710	0	1989	0	23762 SE 2	Maple Val	WA 98038	USA	482000
3	2.5	2430	88426	1	0	0	4	1570	860	1985	0	46611-466	North Ben	WA 98045	USA	452500
4	2	1520	6200	1.5	0	0	3	1520	0	1945	2010	6811 55th	Seattle	WA 98115	USA	640000
3	1.75	1710	7320	1	0	0	3	1710	0	1948	1994	Burke-Gilr	Lake Fore	WA 98155	USA	463000
4	2.5	2920	4000	1.5	0	0	5	1910	1010	1909	1988	3838-4098	Seattle	WA 98105	USA	1400000

- Linear regression algorithm shows **a linear relationship between** a dependent (y) and one or more independent (x) variables, hence called as linear regression.
- Since linear regression shows the linear relationship, which means it **finds how the value** of the dependent variable is **changing** according to the value of the independent variable.

Example:

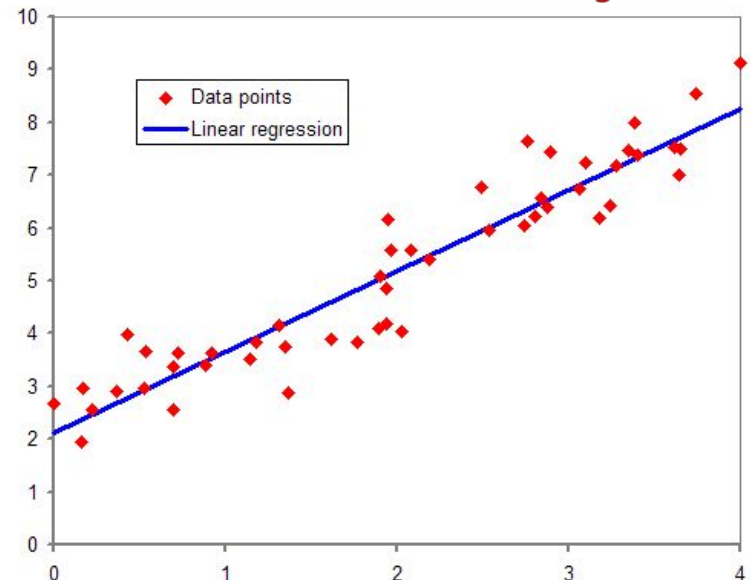
**Height and weight, Year of Experience and salary,
Driving speed and mileage**

- **Regression** takes a group of random variables, thought to be predicting Y, and tries to find a **mathematical relationship** between them.
- This relationship is typically **in the form of a straight line** (linear **regression**) that best approximates all the individual data points



Classification

Regression



Simple Linear Regression

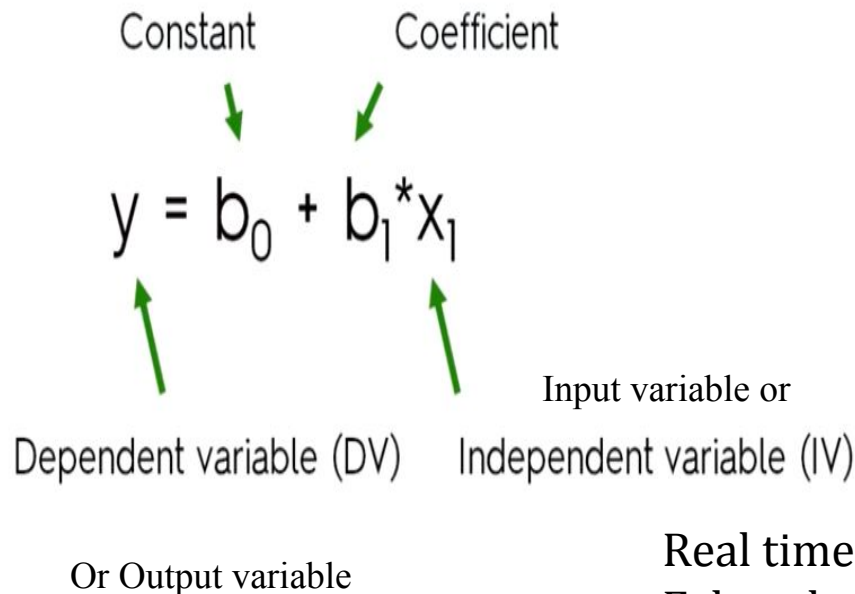
- Simple linear regression is nothing but a straight line with a slop.
- So we use the formula of straight line

Constant Coefficient

$$y = b_0 + b_1 * x_1$$

Dependent variable (DV) Input variable or Independent variable (IV)

Or Output variable



Output is Y
Input is X

$$Y = c + bX \text{ or } Y = mX + c$$

Eg : Fahrenheit = $32 + \frac{9}{5}$ Celsius

$$Y = c + bX$$

Real time Example

$$\text{Fahrenheit} = 32 + (9/5) \text{ Celsius}$$

So

Celsius is independent variable(X)

Fahrenheit is dependent variable (Y)

Three type of correlation between one variable and another

1) Positive correlation

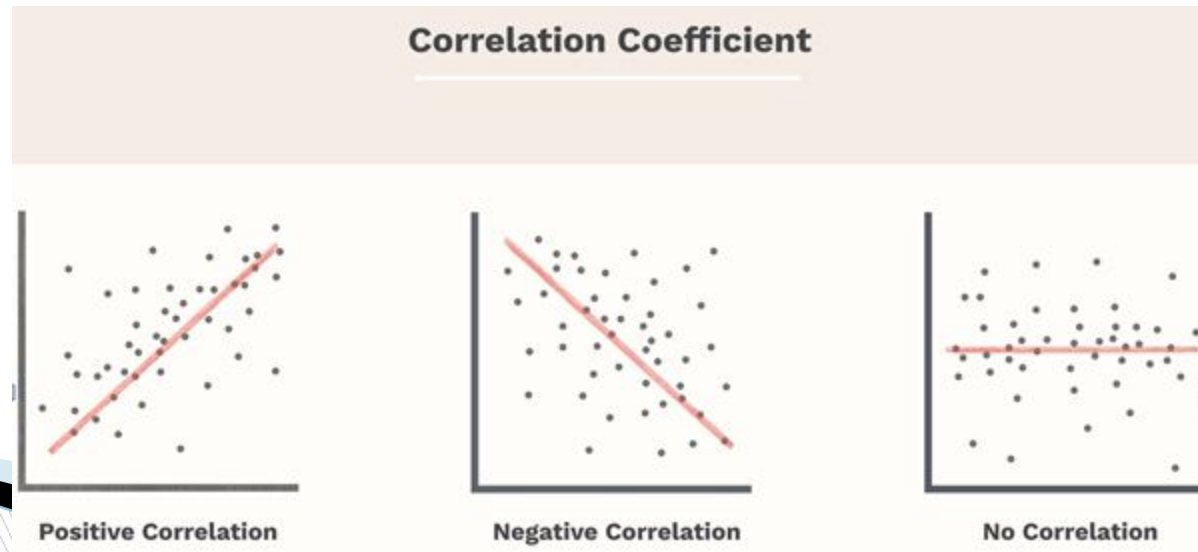
- When input variable x when it increase, output variable y will also increase

2) Negative correlation

- When input variable x increase, output variable Y value will decrease

3) No Correlation

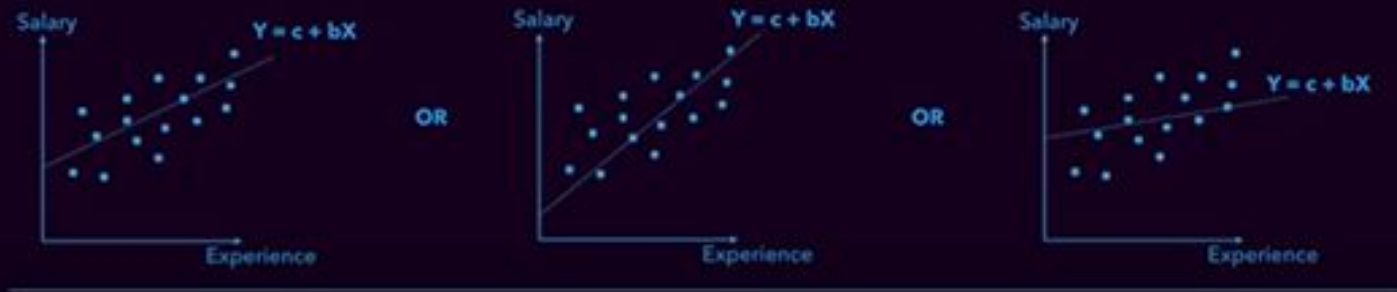
- How much ever we increase the value of X , Output variable Y will be flat and have same value



Finding the best fit line:

- When working with linear regression, our main goal is to find the best fit line that means the error between predicted values and actual values should be minimized.
- The best fit line will have the least error
- The different values for weights or the coefficient of lines (a_0, a_1) gives a different line of regression, so we need to calculate the best values for a_0 and a_1 to find the best fit line, so to calculate this we use cost function

BEST FITTED SLOPE



$$\text{Error} = (\text{Actual}) - (\text{Predicted})$$

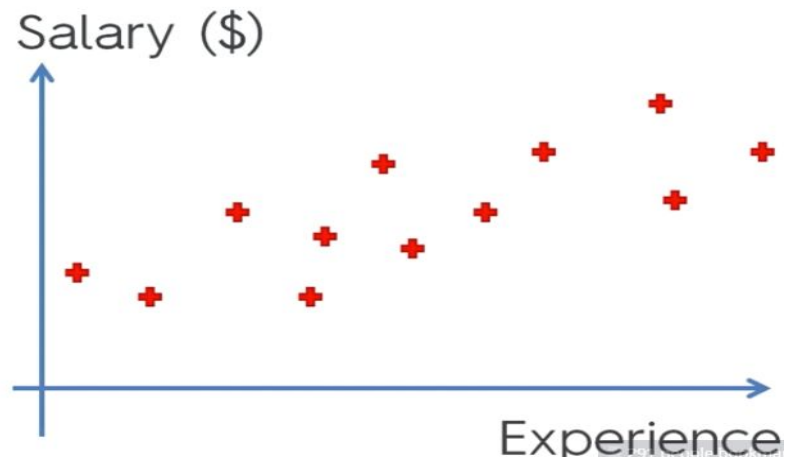
$$\text{Error} = y_1 - \hat{y}_1$$

$$\text{Sum of Squared Error (SSE)} = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

Best fitted slope=Minimum value of SSE

Sample dataset

□ Year of experience vs Salary



A	B
YearsOfExperience	Salary
1.2	38976
1.3	45897
1.5	36987
1.4	40587
1.3	42984
1.7	47986
2	44578
2.2	38789
2.4	46986
2.6	47986
2.9	56642
3	60150
3.2	54445
3.3	58763
3.5	56498
3.9	63218

Linear Regression



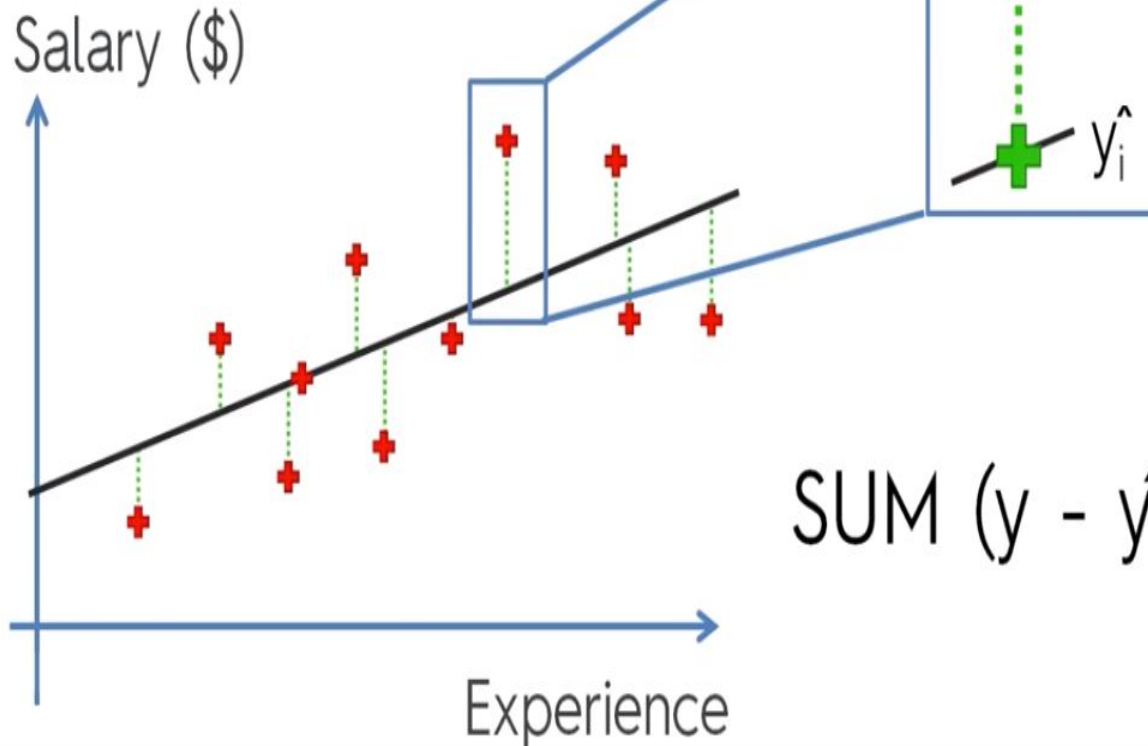
$$y = b_0 + b_1 x$$



$$\text{Salary} = b_0 + b_1 \text{Experience}$$

Linear Regression

Simple Linear Regression:



$$\text{SUM } (y - \hat{y})^2 \rightarrow \min$$

Regression Line

$$y = a_0 + a_1 * x$$

$$a_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

$$a_0 = \bar{y} - a_1 \bar{x}$$

```
x_mean=np.mean(x)
y_mean=np.mean(y)
print("Mean for YearsExperience ",x_mean,"\nMean for Salary ",y_mean)
```

```
Mean for YearsExperience  5.3133333333333335
Mean for Salary  76003.0
```

$$a_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

$$a_0 = \bar{y} - a_1 \bar{x}$$



$$a_1 = \text{sum}((x_i - \text{mean}(x)) * (y_i - \text{mean}(y))) / \text{sum}((x_i - \text{mean}(x))^2)$$

$$a_0 = \text{mean}(y) - a_1 * \text{mean}(x)$$



$$a_0 = y_mean - (a_1 * x_mean)$$



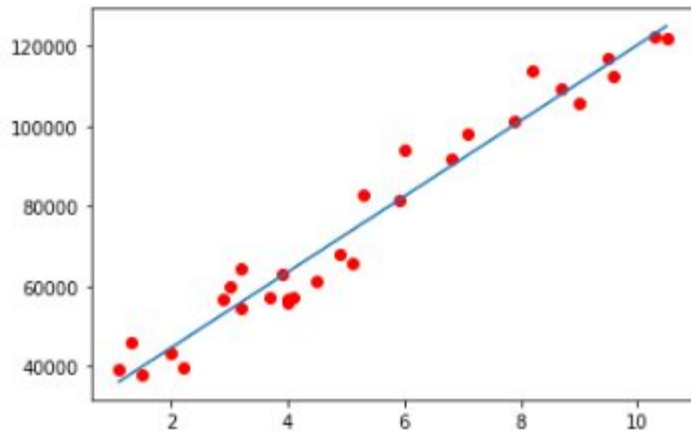
```
num=0
den=0
for i in range(len(x)):
    num+=(x[i]-x_mean)*(y[i]-y_mean)
    den+=(x[i]-x_mean)**2
a1=num/den
a1
```

```
print("Intercept is ",float(a0),"\nslope is ",float(a1))
```

```
Intercept is 25792.20019866869  
slope is 9449.962321455077
```

$$y_{\text{pred}} = a_0 + a_1 * x$$

y_{pred}



Actual value

Predicted value

```
array([[ 36187.15875227],  
       [ 38077.15121656],  
       [ 39967.14368085],  
       [ 44692.12484158],  
       [ 46582.11730587],  
       [ 53197.09093089],  
       [ 54142.08716303],  
       [ 56032.07962732],  
       [ 56032.07962732],  
       [ 60757.06078805],  
       [ 62647.05325234],  
       [ 63592.04948449],  
       [ 63592.04948449],  
       [ 64537.04571663],  
       [ 68317.03064522],  
       [ 72097.0155738 ],  
       [ 73987.00803809],  
       [ 75877.00050238],  
       [ 81546.97789525],  
       [ 82491.9741274 ],  
       [ 90051.94398456],  
       [ 92886.932681  ],  
       [100446.90253816],  
       [103281.8912346 ],  
       [108006.87239533],  
       [110841.86109176],  
       [115566.84225249],  
       [116511.83848464],  
       [123126.81210966],  
       [125016.80457395]])
```

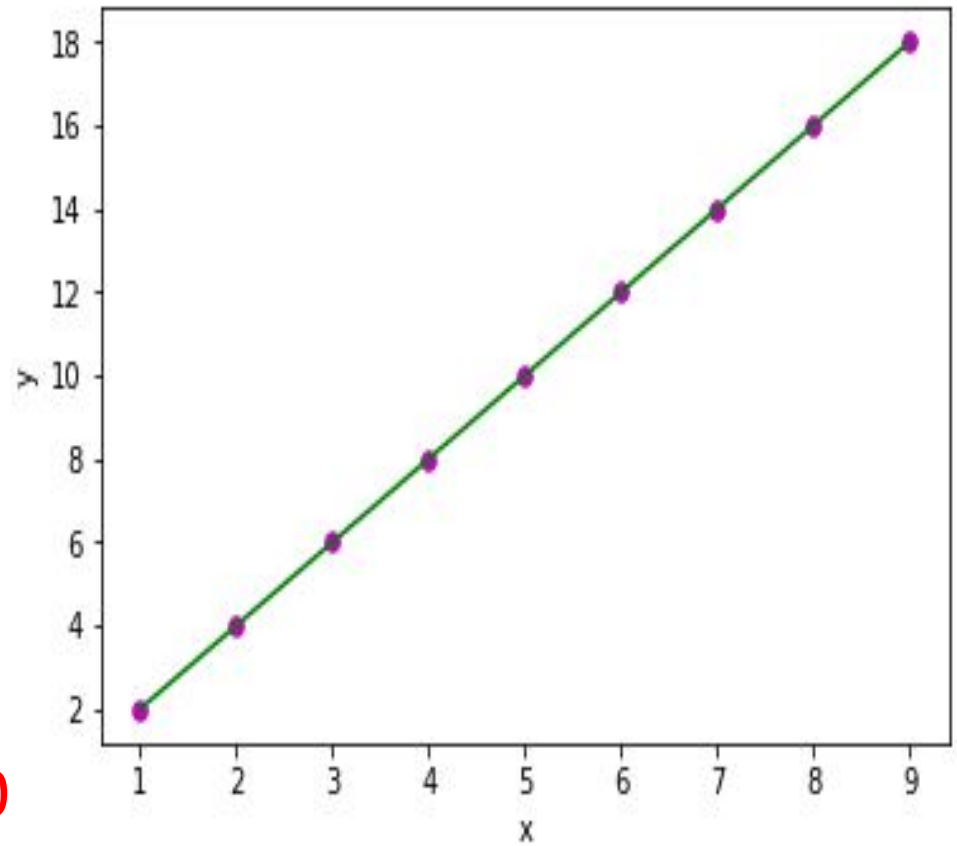
A	B
YearsOfExperience	Salary
1.2	38976
1.3	45897
1.5	36987
1.4	40587
1.3	42984
1.7	47986
2	44578
2.2	38789
2.4	46986
2.6	47986
2.9	56642
3	60150
3.2	54445
3.3	58763
3.5	56498
3.9	63218

$x = ([1, 2, 3, 4, 5, 6, 7, 8, 9])$

$y = ([2, 4, 6, 8, 10, 12, 14, 16, 18])$

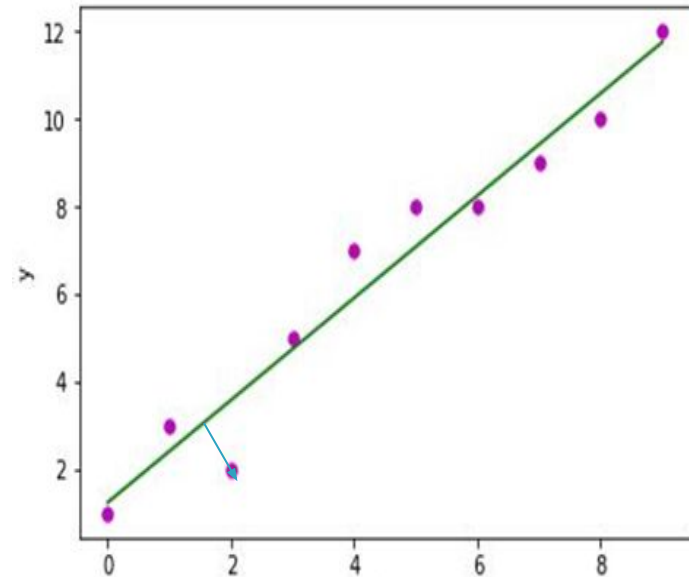
example

- $x = ([1, 2, 3, 4, 5, 6, 7, 8, 9])$
- $y = ([2, 4, 6, 8, 10, 12, 14, 16, 18])$
- $y_{\text{pred}} = a_0 + a_1 * x$
- $a_0 = 0$
- $a_1 = 2$
- $Y = 0 + 2 * X$
- Based on this equ
- $Y_{\text{Pred}} = (2, 4, 6, \dots, 18)$
- $\text{Error} = Y_{\text{given}} - Y_{\text{Pred}} = 0$

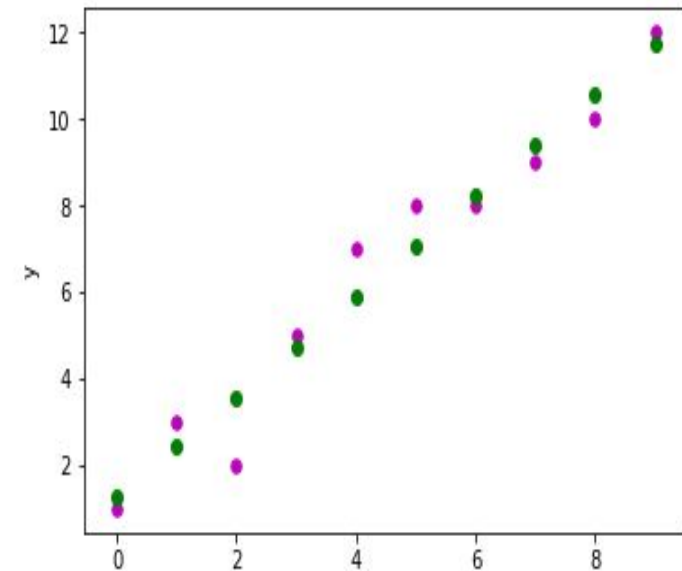


example

- $x = [0, 1, 2, 3, 4, 5, 6, 7, 8, 9]$
- $y = [1, 3, 2, 5, 7, 8, 8, 9, 10, 12]$
- $a_0 + a_1 * x$
-
- $a_0 = 1.2363$
- $a_1 = 1.1696$
- $Y = 1.23 + X * 1.17$
- $Y_{pred} = \{1.23, 2.4\}$
- Error



`plt.plot(x, y_pred, color = "g")`



`plt.scatter(x, y_pred, color = "g")`

Example

i	1	2	3	4	5	6
x	0	0.5	1.0	1.5	2.0	2.5
y	0	1.5	3.0	4.5	6.0	7.5

Result $a_1=3$,
 $a_0=0$

$$y = a_0 + a_1X,$$

X	43	21	25	42	57	59
Y	99	65	79	75	87	81

Result

$$y = a_0 + a_1X,$$

$a_1=0.385$,
 $a_0=65.14$

X	Y
1	3
2	9
3	27
4	64
5	102

example

$$a_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

$$a_0 = \bar{y} - a_1 \bar{x}$$

X	1	2	3	4	5
Y	3	9	27	64	102

	X	Y					Y _{pred}	Error = (Y _{pred} - Y) ²
	1	3	-2	4	-38	76	-9.6	158.76
	2	9	-1	1	-32	32	15.7	44.89
	3	27	0	0	-14	0	4.1	196
	4	64	1	1	23	23	66.3	5.29
	5	102	2	4	61	122	91.6	108.16
Sum	15	205		10		253		513.1
Mean								

$$a_1 = 253/10 = 25.3$$

$$a_0 = -34.9$$

The equation can also be written as

$$Y = a_0 + a_1 X + \text{Residual Error}$$

$$Y = a_0 + a_1 X + (Y - Y_{\text{pred}})$$

Ex : For X=1, Y=3, Y_{pred} = -9.6

$$Y = -34.9 + (25.3 \times 1) + (3 - (-9.6)) = 3$$

Performance evaluation

1) R-squared method or Coefficient of determination (r^2)

- R Square is calculated by the sum of squared of prediction error divided by the total sum of the square which replaces the calculated prediction with mean.
- R Square value is between 0 to 1 and a bigger value indicates a better fit between prediction and actual value.
- The lowest possible value of r^2 is 0 and the highest possible value is 1. Put simply, the better a model is at making predictions, the closer its r^2 will be to 1.

$$\text{R-squared} = 1 - (\text{SSR} / \text{SST})$$

Where, SSR = sum of squared errors $\sum (y_{\text{actual}} - y_{\text{predicted}})^2$

SST = sum of squares $\sum (y_{\text{actual}} - y_{\text{mean}})^2$

R-squared

```
num_r=0
den_r=0
for i in range(len(x)):
    num_r+=(y[i]-y_pred[i])**2
    den_r+=(y[i]-y_mean)**2
r_sq=1-(num_r/den_r)
r_sq
```

```
print("The value of coefficient of determination R_square is ",float(r_sq))
```

```
The value of coefficient of determination R_square is  0.9569566641435086
```

Mean Square Error

- ❑ MSE is calculated by the sum of square of prediction error which is real output minus predicted output and then divide by the number of data points.
- ❑ It gives you an absolute number on how much your predicted results deviate from the actual number. You cannot interpret many insights from one single result but it gives you a real number to compare against other model results and help you select the best regression model.

$$MSE = \frac{1}{n} \sum \left(\underbrace{y - \hat{y}}_{\substack{\text{The square of the difference} \\ \text{between actual and} \\ \text{predicted}}} \right)^2$$

Root Mean Squared Error (RMSE)

- Root Mean Square Error(RMSE) is the square root of MSE. It is used more commonly than MSE because firstly sometimes MSE value can be too big to compare easily.
- Secondly, MSE is calculated by the square of error, and thus square root brings it back to the same level of prediction error and makes it easier for interpretation.

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (\hat{y}_i - y_i)^2}$$

Mean Absolute Error

- Mean Absolute Error(MAE) is similar to Mean Square Error(MSE). However, instead of the sum of square of error in MSE, MAE is taking the sum of the absolute value of error.
- Compare to MSE or RMSE, MAE is a more direct representation of sum of error terms. **MSE gives larger penalization to big prediction error by square it while MAE treats all errors the same.**

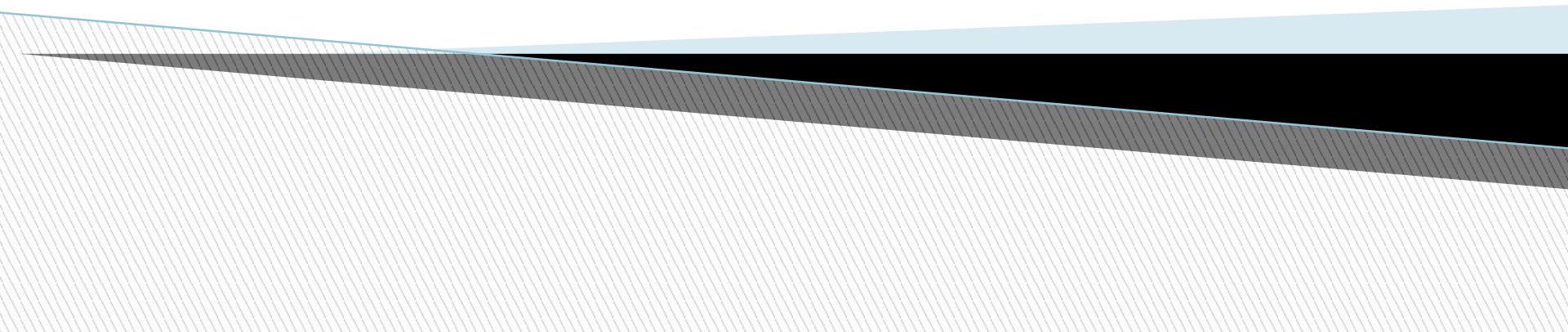
The diagram illustrates the Mean Absolute Error (MAE) formula with the following components and annotations:

- Divide by the total number of data points:** Points to the $\frac{1}{n}$ term in the formula.
- Sum of:** Points to the summation symbol Σ .
- Actual output value:** Points to the y term inside the absolute value.
- Predicted output value:** Points to the \hat{y} term inside the absolute value.
- The absolute value of the residual:** Points to the entire absolute value expression $|y - \hat{y}|$.

$$MAE = \frac{1}{n} \sum |y - \hat{y}|$$

- R Square/Adjusted R Square is better used to explain the model to other people because you can explain the number as a percentage of the output variability.
- MSE, RMSE, or MAE are better be used to compare performance between different regression models.

Multiple Linear Regression



- Multiple linear regression is a method we can use to quantify the relationship between two or more predictor variables and a response variable.

Consider the following dataset with one response variable y and two predictor variables X_1 and X_2 :

y	X_1	X_2
140	60	22
155	62	25
159	67	24
179	70	20
192	71	15
200	72	14
212	75	14
215	78	11

Calculate X_1^2 , X_2^2 , X_1y , X_2y and

y	X_1	X_2
140	60	22
155	62	25
159	67	24
179	70	20
192	71	15
200	72	14
212	75	14
215	78	11
Mean	181.5	18.125
Sum	1452	145

X_1^2	X_2^2	X_1y	X_2y	X_1X_2
3600	484	8400	3080	1320
3844	625	9610	3875	1550
4489	576	10653	3816	1608
4900	400	12530	3580	1400
5041	225	13632	2880	1065
5184	196	14400	2800	1008
5625	196	15900	2968	1050
6084	121	16770	2365	858
Sum	38767	101895	25364	9859

$$\Sigma X_1^2 = \Sigma X_1^2 - (\Sigma X_1)^2 / n = 38,767 - (555)^2 / 8 = \mathbf{263.875}$$

$$\Sigma X_2^2 = \Sigma X_2^2 - (\Sigma X_2)^2 / n = 2,823 - (145)^2 / 8 = \mathbf{194.875}$$

$$\Sigma X_1y = \Sigma X_1y - (\Sigma X_1 \Sigma y) / n = 101,895 - (555 * 1,452) / 8 = \mathbf{1,162.5}$$

$$\Sigma X_2y = \Sigma X_2y - (\Sigma X_2 \Sigma y) / n = 25,364 - (145 * 1,452) / 8 = \mathbf{-953.5}$$

$$\Sigma X_1X_2 = \Sigma X_1X_2 - (\Sigma X_1 \Sigma X_2) / n = 9,859 - (555 * 145) / 8 = \mathbf{-200.375}$$

Calculate b_0 , b_1 , and b_2 .

$$b_1 = \frac{(\sum x_2^2)(\sum x_1 y) - (\sum x_1 x_2)(\sum x_2 y)}{(\sum x_1^2)(\sum x_2^2) - (\sum x_1 x_2)^2}$$

and

$$b_2 = \frac{(\sum x_1^2)(\sum x_2 y) - (\sum x_1 x_2)(\sum x_1 y)}{(\sum x_1^2)(\sum x_2^2) - (\sum x_1 x_2)^2}$$

$$b_1 = [(194.875)(1162.5) - (-200.375)(-953.5)] / [(263.875)(194.875) - (-200.375)^2] = \mathbf{3.148}$$

$$b_2 = [(263.875)(-953.5) - (-200.375)(1152.5)] / [(263.875)(194.875) - (-200.375)^2] = \mathbf{-1.656}$$

$$b_0 = \bar{Y} - b_1 \bar{X}_1 - b_2 \bar{X}_2$$

$$b_0 = 181.5 - 3.148(69.375) - (-1.656)(18.125) = \mathbf{-6.867}$$

Multiple Linear Regression Equation

$$\hat{y} = b_0 + b_1 * x_1 + b_2 * x_2$$

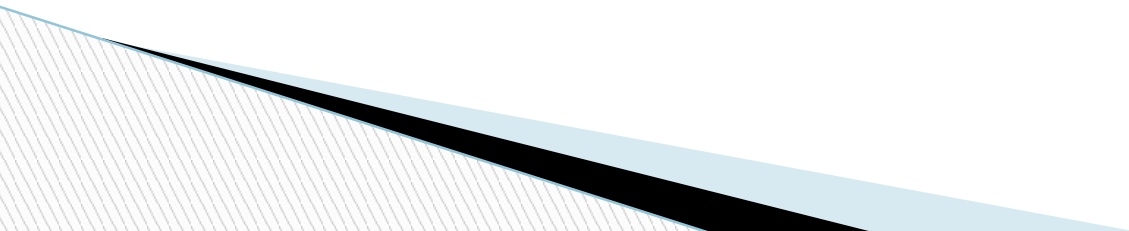
$$\hat{y} = -6.867 + 3.148x_1 - 1.656x_2$$

$b_0 = -6.867$. When both predictor variables are equal to zero, the mean value for y is -6.867 .

$b_1 = 3.148$. A one unit increase in x_1 is associated with a 3.148 unit increase in y , on average, assuming x_2 is held constant.

$b_2 = -1.656$. A one unit increase in x_2 is associated with a 1.656 unit decrease in y , on average, assuming x_1 is held constant.

what is polynomial regression



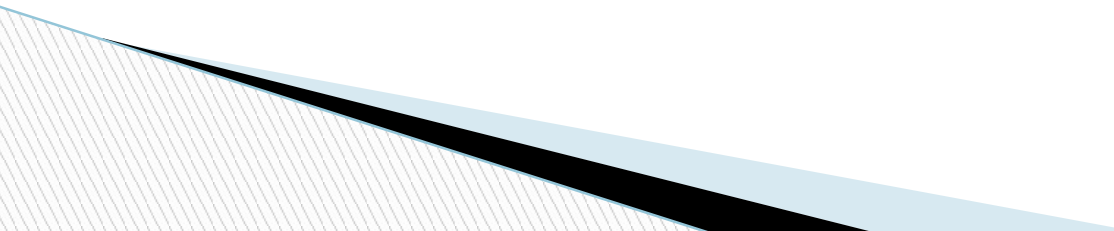
Polynomial regression

- In simple linear regression, the relationship between the independent variable and the dependent variable is modeled as a straight line.
- However, in polynomial regression, the relationship can be modeled as a curve, allowing for more flexibility in capturing non-linear relationships between variables.

The polynomial regression model can be expressed as:

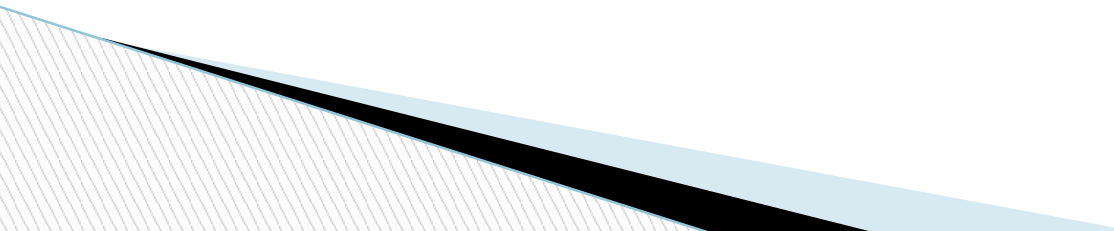
$$y = b_0 + b_1x + b_2x^2 + b_3x^3 + \dots + b_nx^n$$

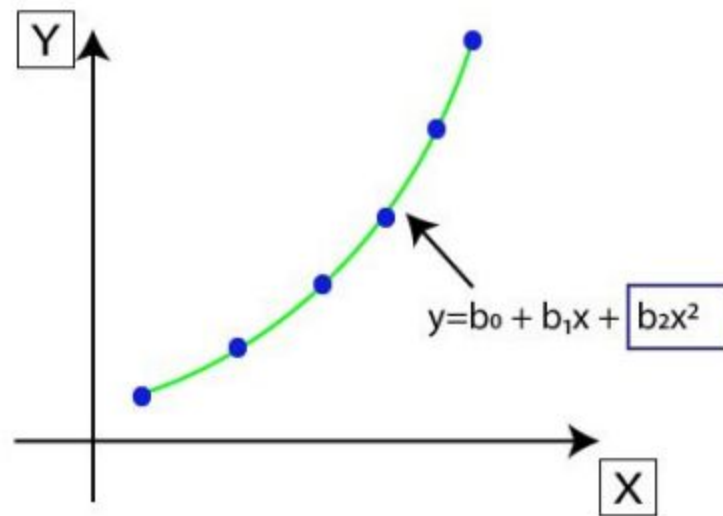
where y is the dependent variable, x is the independent variable, and $b_0, b_1, b_2, \dots, b_n$ are the coefficients that determine the shape of the polynomial curve.



- By adjusting the values of n (the degree of the polynomial), you can create different polynomial regression models.

For example:

- Linear regression: $n=1$
 - Quadratic regression: $n=2$
 - Cubic regression: $n=3$
 - Higher-degree polynomial regression: $n>3$
-
- Polynomial regression can be useful when the relationship between the variables is not linear and cannot be adequately captured by a straight line.
 - It allows for more complex relationships to be modeled, such as quadratic, cubic, or higher-order polynomial relationships.
- 



Let the quadratic polynomial regression model be

$$y = a_0 + a_1 * x + a_2 * x^2$$

The values of **a_0** , **a_1** and **a_2** are calculated using the following system of equations:

$$\begin{aligned}\sum y_i &= na_0 + a_1(\sum x_i) + a_2(\sum x_i^2) \\ \sum y_i x_i &= a_0(\sum x_i) + a_1(\sum x_i^2) + a_2(\sum x_i^3) \\ \sum y_i x_i^2 &= a_0(\sum x_i^2) + a_1(\sum x_i^3) + a_2(\sum x_i^4)\end{aligned}$$

Example

X	Y
3	2.5
4	3.2
5	3.8
6	6.5
7	11.5

X	Y
3	2.5
4	3.2
5	3.8
6	6.5
7	11.5

	x	y	x ²	x ³	x ⁴	yx	yx ²
	3	2.5	9	27	81	7.5	22.5
	4	3.2	16	64	256	12.8	51.2
	5	3.8	25	125	625	19	95
	6	6.5	36	216	1296	39	234
	7	12	49	343	2401	80.5	563.5
Σ	25	27.5	135	775	4659	158.8	966.2

$$\sum y_i = na_0 + a_1(\sum x_i) + a_2(\sum x_i^2)$$

$$\sum y_i x_i = a_0(\sum x_i) + a_1(\sum x_i^2) + a_2(\sum x_i^3)$$

$$\sum y_i x_i^2 = a_0(\sum x_i^2) + a_1(\sum x_i^3) + a_2(\sum x_i^4)$$

Using the given data we

$$27.5 = 5a_0 + 25a_1 + 135a_2$$

$$158.8 = 25a_0 + 135a_1 + 775a_2$$

$$966.2 = 135a_0 + 775a_1 + 4659a_2$$

Solving this system of equations we get

$$a_0 = 12.4285714$$

$$a_1 = -5.5128571$$

$$a_2 = 0.7642857$$

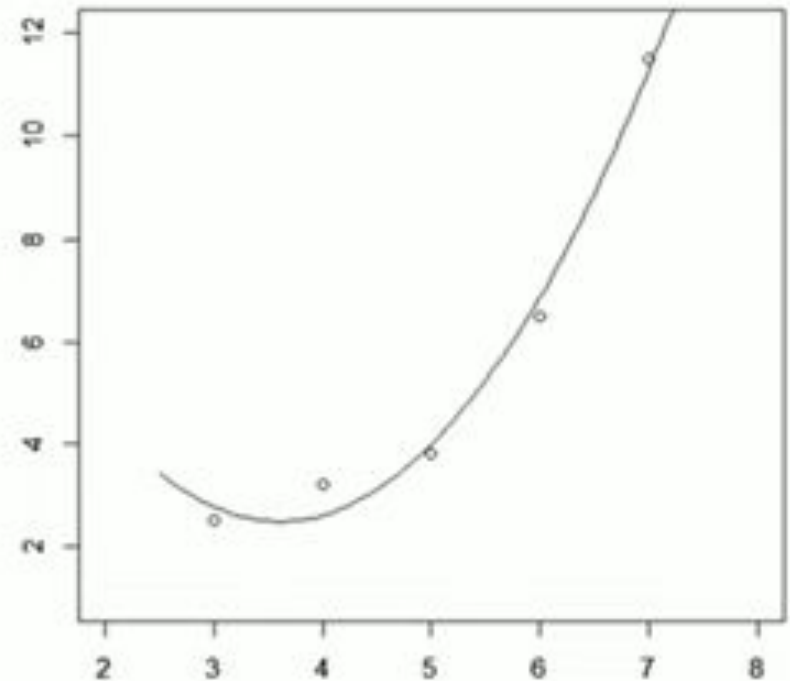
$$a_0 = 12.4285714$$

$$a_1 = -5.5128571$$

$$a_2 = 0.7642857$$

The required quadratic polynomial model is

$$y = 12.4285714 - 5.5128571x + 0.7642857x^2$$



Changing the value of x 0 to n we get the polynomial curve

If u know the value of x we can calculate the value of y