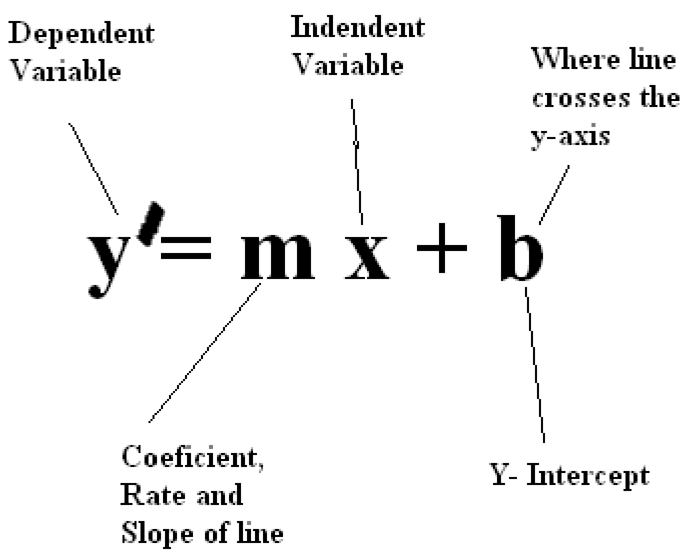
Regression means=finding an equation of line that represents relationship b/w dependent and independent variables.

Simple Regression:

Years Experience(x)	Salary(y)
1	10000
3	22000
5	32000
6	?

Now we need to find a regression line(Predicted y) :



Here y dash represents predicted output for each x

$$m = \frac{\overline{x} \cdot \overline{y} - \overline{xy}}{(\overline{x})^2 - \overline{x^2}}$$

Here:

X bar=Mean of X

Y bar=Mean of y

XY bar=Mean of x.y

$$b = \overline{y} - m\overline{x}$$

Years Experience(x)	Salary(y)	x.y	X sqr		
1	10000	10000	1		
3	22000	66000	9		
5	32000	160000	25		
X bar=3	y bar=21333	xy bar=78666	X sqr bar=11.6		

M=(3*21333-78666)/(9-11.6)

M=5641.15

B=21333-5641.5*3

21333-16924.5

B=4408.5

Now, value of y at x=6

Y dash=5641*6+4408.5

=38254

Multiple Regressions:

If we have k independent variables and a slope for each.

The prediction equation is:

$$Y' = a + b_1 X_1 + b_2 X_2 + \ldots + b_k X_k$$

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

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

$$a = \overline{Y} - b_1 \overline{X}_1 - b_2 \overline{X}_2$$

	x1	**x2**	**y**	
0	2	0	4	
1	3	1	8	
2	1	1	2	
3	3	0	5	
4	4	1	?	

Now,

	x1	**x2**	**y**	**X1 sqr**	**x2 sqr**	**x1.y**	**x2.y**	X1.x2
0	2	0	4	4	0	8	0	0
1	3	1	8	9	1	24	8	3
2	1	1	2	1	1	2	2	1
3	3	0	5	9	0	15	0	0
Sum	**9**	**2**	**19**	**23**	**2**	**49**	**10**	**4**
Mean(bar)	2.25	0.5	4.75	5.75	0.5	12.25	2.5	1

B2= (23x10)- (4x49)/(23x2)-16

= (230-196)/(46-16)

B2= 1.13

B1= (2x 49)-(4x10)/(23x2)-16

=98-40/46-16

=58/30

B1=1.93

A=4.75-(1.93x2.25)-(1.13x.5)

=4.75-4.34-.565

A=-.155

Now , value of y at x1=4 and x2=1:

Y=-.155+1.93x4+1.13x1

=.975+7.72

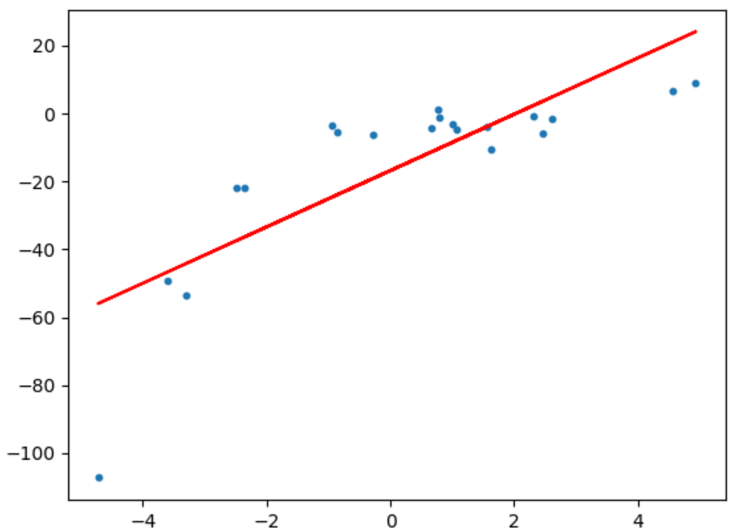
Y=8.695

Polynomial Regression:

If regression line

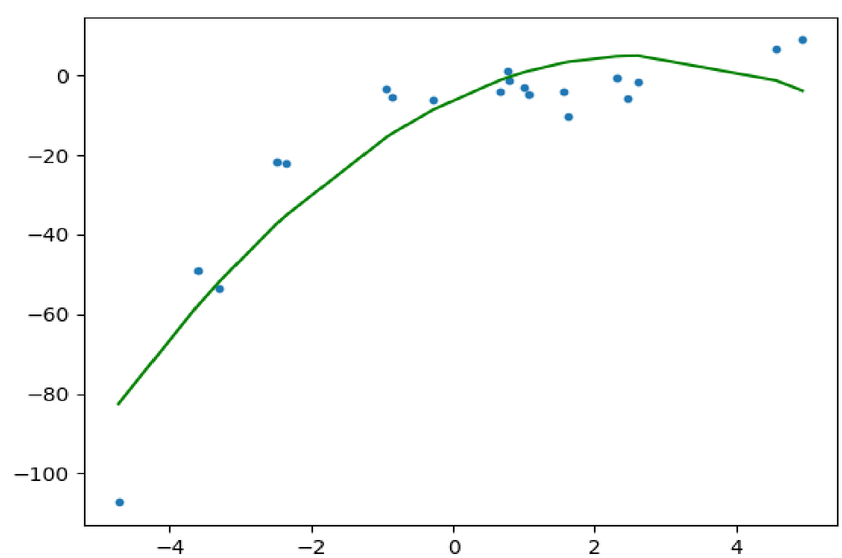
Y=b+mx

does not capture most of the actual data points like



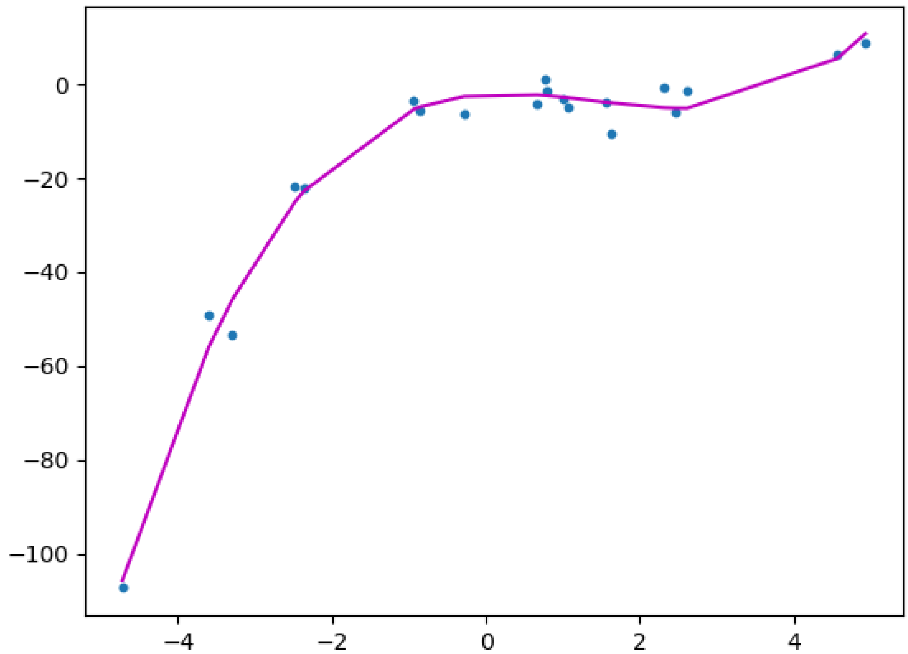
Then we should use polynomial regression using following equation

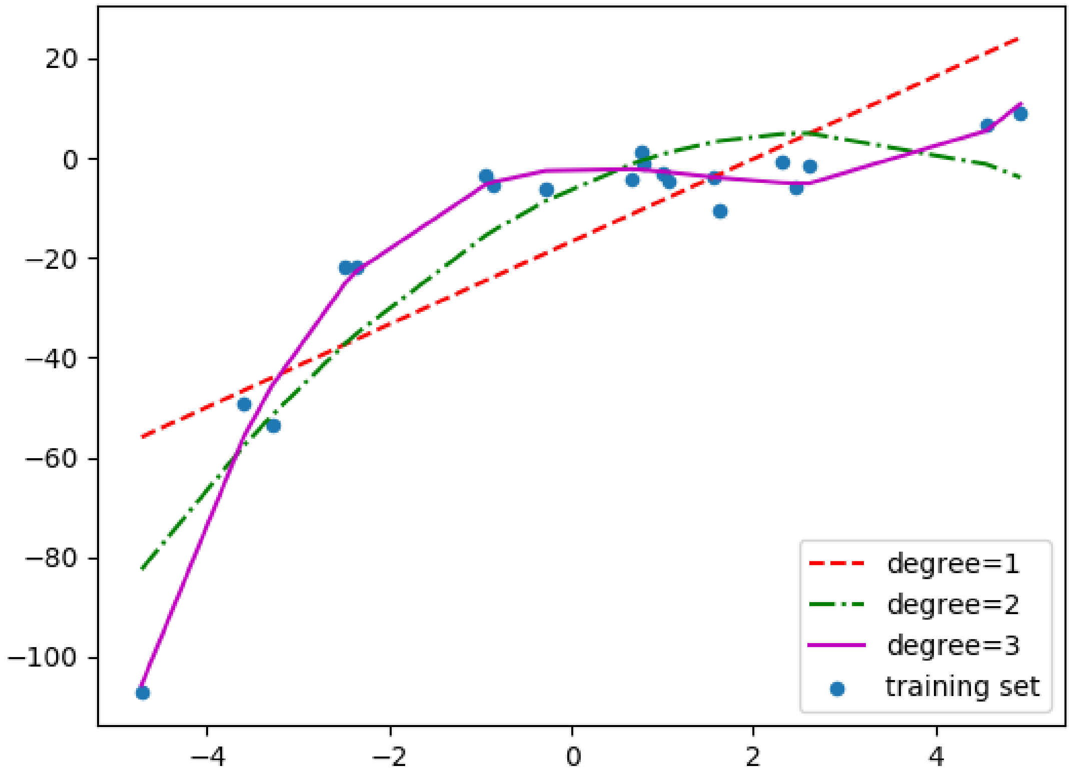
Y=b+m1x+m2x(sqr)   (polynomial with 2 order)



We may also increase order

Y=c+m1x+m2x(sqr)+m3x(cube) 3 order





Now Most Challenging part of polynomial regression is to find out values of b, m1, m2.

$$\begin{bmatrix} n & \sum x & \sum x^2 \\ \sum x & \sum x^2 & \sum x^3 \\ \sum x^2 & \sum x^3 & \sum x^4 \end{bmatrix} \begin{bmatrix} b \\ m_1 \\ m_2 \end{bmatrix} = \begin{bmatrix} \sum y \\ \sum xy \\ \sum x^2 y \end{bmatrix}$$

here n = No. of samples

Consider Following dataset:

Level(X)	Salary(Y)
1	4.5
2	5.0
3	6.0
4	8.0
5	11.0
6	15.0
7	20.0
8	30.0
9	50.0
10	100.0

Here n=10

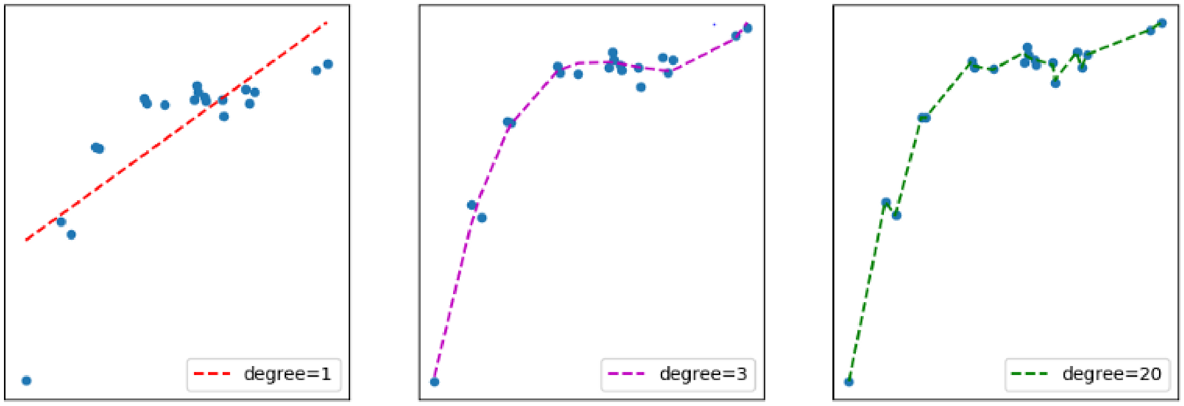
self

The Bias vs Variance trade-off

Bias refers to the error due to the model's simplistic assumptions in fitting the data. A high bias means that the model is unable to capture the patterns in the data and this results in **under-fitting**.

Variance refers to the error due to the complex model trying to fit the data. High variance means the model passes through most of the data points and it results in **over-fitting** the data.

The below picture summarizes our learning.



Underfit
High Bias
Low Variance

Correct Fit
Low Bias
Low Variance

Overfit
Low Bias
High Variance