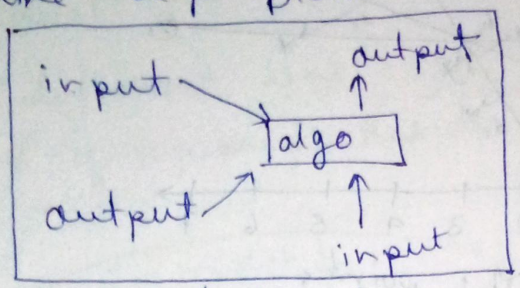
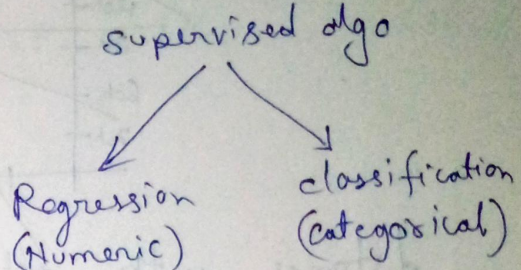
01 Oct 2023

Linear Regression

It is algorithm that is used to solve supervised machine ml problem.

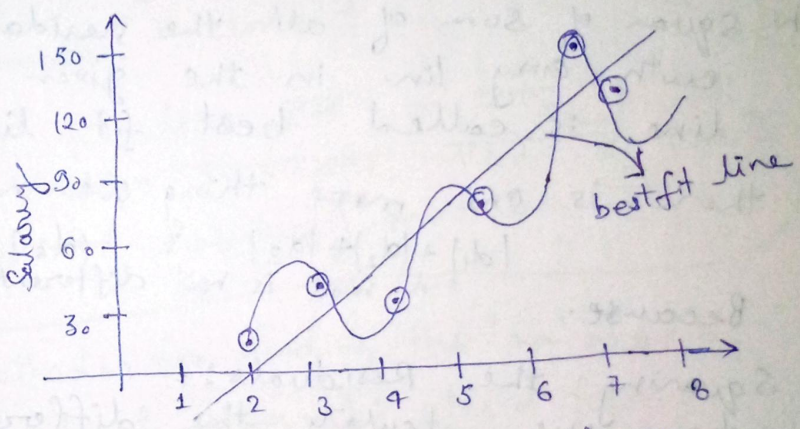


Supervised learning

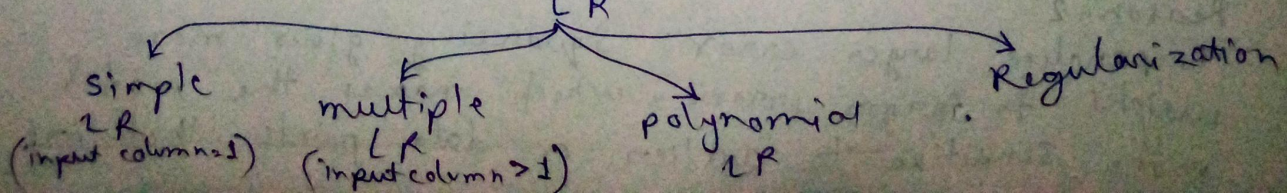


experience	Salary
5.6	75000
7.2	120000
3.1	47000
4.4	37000
2.0	16000
⋮	⋮
7	140000

training data

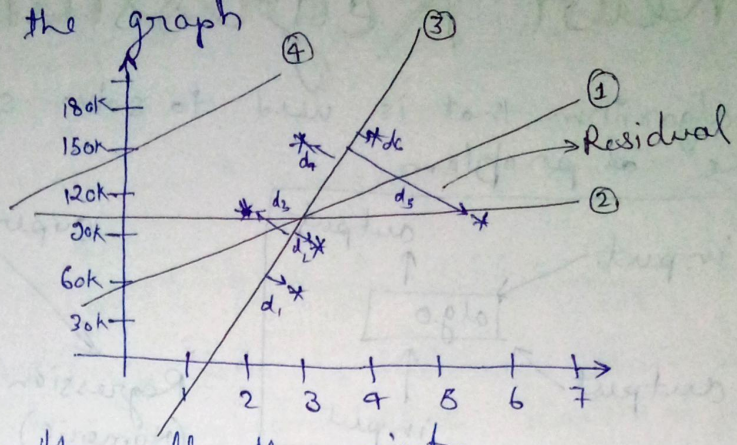


* machine learning Algorithm का use करके हम एक ऐसा model तैयार करते है जो भले ही वह Training data पर accurate work ना करे लेकिन वह Testing data पर accurately work करना चाहिए
* The curve may be perfectly work with training data but it not perfectly work with testing data hence our end goal is to make the model which is perfectly work with testing data provided by user.



How to select bestfit line?

The best fit line is always most closer to each point on the graph



distance with all the points

$d_1^2 + d_2^2 + d_3^2 + \dots \quad d_6^2 = \min(D)$

*the squar of sum of all the residal must be minimum with any line in the given then that ~~that~~ line is called best fit line

* there is one more thing we not use mod $|d_1|$

$|d_1| + |d_2| + |d_3| \dots \quad (d_6) = \min(D)$

and mod is not differentiable at zero

Because.

Squaring the Residuals:

when we calculate the difference between the observed Y-values and the predicted Y-values (residuals). we square these differences for this Resason.

Reason-1

To. Remove the sign: Squaring ensures that all residuals are positive. This is important because. if we only added up the differences without squaring them, the positive and negative residuals would cancel each other out, and we wouldn't be able to assess the overall goodness of fit.

Reason-2

To penalize larger error: Squareing gives more weight to larger errors, which makes the model more sensitive to outlier or data points that are

that have a significant impact on the overall fit of the line.

$$d_1^2 + d_2^2 + d_3^2 \cdots d_n^2 = D$$

In the above D is called sum of squared Residual.

for different line SSR_1 SSR_2 SSR_3 --- SSR_n
(sum of squared Residual)

$y = mx + b$

- m → slop of line (angular movement)
- b → y-intercept movement

for finding best fit line हम m और b का एक ऐसा combination निकालना चाहते है जिसमे SSR (sum of squared Residual) minimum हो।

there is two methods for finding the 'm' and 'b'

(i) OLS (ordinary least square): closed form
it is set of formula (statistical method)

(ii) Gradient Descent (calculus method):
Non closed form. also called approximation based technique, it is use full for higher dimension dataset. SGDRegression class.
Hence we use.

Regression Intution:

for a 2D line: $y = mx + b$

for a 3D line: $y = ax_1 + bx_2 + c$

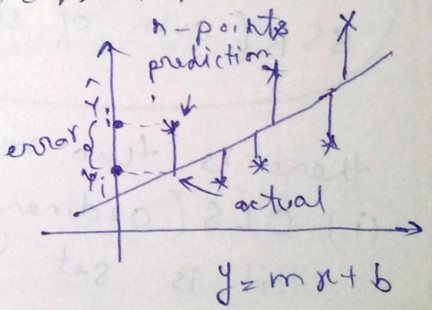
for a 4D line $y = ax_1 + bx_2 + cx_3 + d$ → offset

x_1	x_2	x_3	y
Exp	profile	Gender	salary
7.5	Web	M	2.1
2.5	Android	M	60k
6	ML	F	1.8 L
8	Web	F	2.5k
2.5	ML	F	80k

* in the above. a, b, c is weight
* यह value y calculate करने में a, b, c का क्या weight होगा it means अगर a और y में strong relation होगा तब a का weight ज्यादा होगा।

Ordinary least Squares Regression.

$$m = \frac{\sum_{i=0}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=0}^{n} (x_i - \bar{x})^2}$$



when $\bar{x}, \bar{y}$ = mean value.

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

proof:

sum of square $= (y_1 - \hat{y}_1)^2 + (y_2 - \hat{y}_2) \cdots + (y_n - \hat{y}_n)$

$$E = SS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \text{error function}$$

$$\because \hat{y}_i = mx_i + b$$

$$E(m, b) = SS = \sum_{i=1}^{n} (y_i - (mx_i + b))^2$$

the error function depends on the two values m, b

in the above equation we have to calculate partial differentiation w.r.t. m and b minimization.

$$\frac{d}{db}\sum\left(y_i-(mx_i+b)\right)^2=0 \quad \text{——— (i)}$$

$$\frac{d}{dm}\sum\left(y_i-(mx_i+b)\right)^2=0 \quad \text{——— (ii)}$$

for equation (i)

$$\frac{\partial F}{\partial b}=\sum\frac{d}{db}\left(y_i-(mx_i+b)\right)^2=0$$

$$=\sum 2\left(y_i-(mx_i+b)\right)\frac{d}{db}\left(y_i-(mx_i+b)\right) \quad \text{——— chain rule}$$

$$=\sum -2\left(y_i-(mx_i+b)\right)=0$$

$$=\sum\left(y_i-(mx_i+b)\right)=0$$

$$=\sum\left(y_i-mx_i-b\right)=0$$

$$=\sum y_i-\sum mx_i-\sum b=0$$

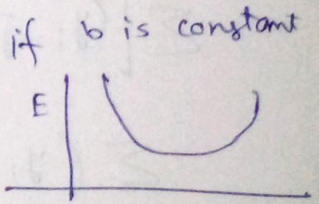
$$=\sum y_i-m\sum x_i-nb=0$$

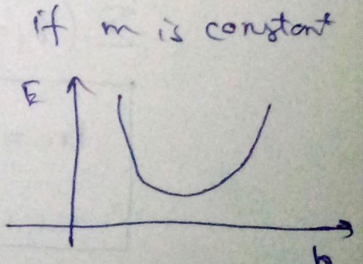
$$=\frac{\sum y_i}{n}-\frac{m}{n}\sum x_i-b=0$$

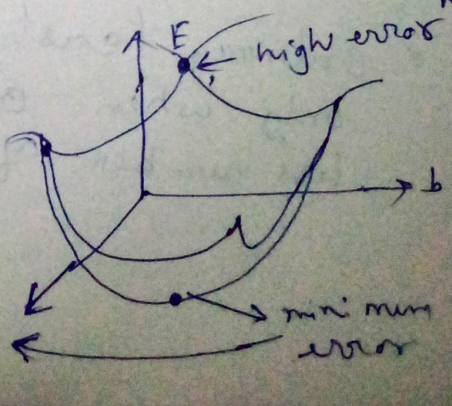
$$=\bar{y}-m\bar{x}-b=0$$

$$=\boxed{b=\bar{y}-m\bar{x}}$$

or $b=\dfrac{\sum Y-m\sum X}{n}$







for equation (ii)

$$\frac{\partial E}{\partial m} = \frac{d}{dm} \Sigma \left(y_i - (mx_i + b)\right)^2 = 0$$

$$\Sigma \frac{d}{dm} \left(y_i - (mx_i + \bar{y} - m\bar{x})\right)^2 = 0$$

~~$\Sigma 2 \left(y_i - (mx_i + \bar{y} - m\bar{x})\right)$~~

$$\Sigma 2 \left(y_i - (mx_i + \bar{y} - m\bar{x})\right) - (x_i - \bar{x}) = 0 \quad \text{— by chain rule}$$

$$-2\Sigma \left(y_i - (mx_i + \bar{y} - m\bar{x})\right)(x_i - \bar{x}) = 0$$

$$\Sigma \left(y_i - (mx_i + \bar{y} - m\bar{x})\right)(x_i - \bar{x}) = 0$$

$$\Sigma \left[(y_i - \bar{y}) - m(x_i - \bar{x})\right](x_i - \bar{x}) = 0$$

$$\Sigma \left[(y_i - \bar{y})(x_i - \bar{x}) - m(x_i - \bar{x})^2\right] = 0$$

$$\Sigma (y_i - \bar{y})(x_i - \bar{x}) - m \Sigma (x_i - \bar{x})^2 = 0$$

$$\Sigma (y_i - \bar{y})(x_i - \bar{x}) = m \Sigma (x_i - \bar{x})^2$$

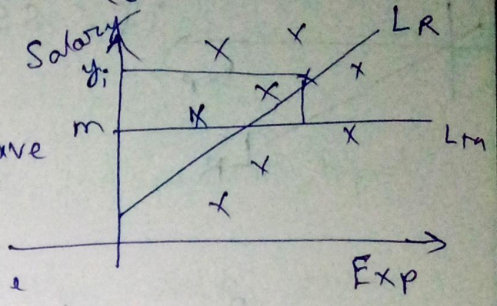
$$\boxed{m = \frac{\Sigma (y_i - \bar{y})(x_i - \bar{x})}{\Sigma (x_i - \bar{x})^2}} \quad \text{or} \quad \frac{n\Sigma XY - \Sigma X \Sigma Y}{n\Sigma X^2 - (\Sigma X)^2}$$

* Ordinary least squares Regression is used only when our data set is small and less number of columns.

Regression Matrics:

(i) R^2 – coefficient of Determination (goodness of fit)

(ii) Adjusted R^2



* L_M line is used when we have Not a experience column.
* L_R line is used when we have two columns Salary and Experience.
* R^2 is used to find How much better L_R from L_M

$$R^2 = \frac{Var_M - Var_R}{Var_M}$$

$$R^2 = 1 - \frac{Var_R}{Var_M} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_M)^2}$$

$$Var_M = SSR_M = \sum_{i=1}^{n}(y_i - \hat{y}_M)^2$$

$$Var_R = SSR_R = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

in the above $0 < R^2 < 1$

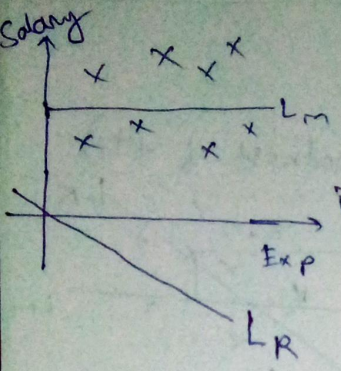
R^2 is calculated in % example 80%.

80% of R^2 means. salary and experience के correlation का used करके हम we are only able to explain the 80% variesen in data.

Number of column $\propto R^2$

interview question - क्या R^2 -Ve हो सकता है?

Answer – Yes



$$R^2 = 1 - \frac{SSR_R}{SSR_m}$$

if $SSR_R > SSR_m$ R^2 is -Ve

$$R^2_{adj} = 1 - \left[\frac{(1-R^2)(n-1)}{(n-k-1)}\right]$$

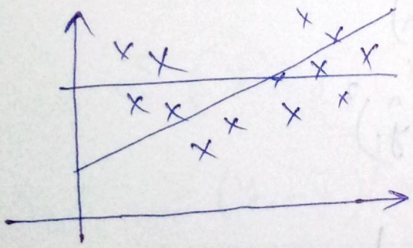
n = Number of raw in data.

k = Number of ~~input~~ independent column

Polynomial Regression:

$y = \beta_0 + \beta_1 x$ → simple linear Regression.

$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots \beta_n x_n$ → Multiple linear Regression



x	y
35	100

degree 2

$x^0 = 1$, $x^1 = 35$, $x^2 = 1225$, Y

Hence $y = \beta_0 + \beta_1 x + \beta_2 x^2$

* if degree is very high then overfitting
* if degree is very low then underfitting

x_1	x_2	y

input column is two for degree = 2

$y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_2 + \beta_4 x_2^2$

5. Adjusted R2 score:

$$R2_{adjusted} = 1 - \left[\frac{(1 - R_2)(n-1)}{(n-1-k)}\right]$$

n = number of rows.

k = total number of independent input column

* if irrelevant R2 score is added then Adj R_2 score is decrease
* if relevant column is added then adj R_2 score is increase.
* adjusted R2 score is usefull when we have a multiple linear regression.

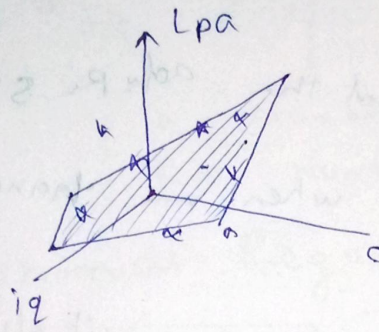
interview question

Ques: between R^2 and adj R^2 which will always be bigger

Multiple linear Regression

- it have more than one independent columns.
- example

cgpa	gender	iq	package (Lpa)
8	Male	124	3.8
⋮	⋮	⋮	⋮



there we draw the plane.
also called best fit plane
this is for 3-D.
for the 4-D hyperplane.

for 2D Data : $y = mx + b$

for 3D Data : $y = m x_1 + m x_2 + b$

OR

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

for 4D Data: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$

for nD Data : $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$

$$y = \beta_0 + \sum_{i=1}^{n} \beta_i x_i$$

here coefficient $\beta_0, \beta_1, \beta_2 \cdots \beta_n$ are weights
e.g. y calculate करने में x_2 का कितना weightage है,

$$lpa = \beta_0 + \beta_1 cgpa + \beta_2 \times iq$$

if $\beta_1 > \beta_2$

cgpa has more weightage than iq
for the calculating lpa

β_0 = offset / intercept value

cgpa	iq	gender	lpa
x_{11}	x_{12}	x_{13}	y → actual

predicted

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

for calculating $\hat{y}$ (predicted)

$$\begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_{100} \end{bmatrix} = \begin{bmatrix} \beta_0 & \beta_1 x_{11} & \beta_2 x_{12} & \beta_3 x_{13} \\ \beta_0 & \beta_1 x_{21} & \beta_2 x_{22} & \beta_3 x_{23} \\ \vdots & \vdots & \vdots & \vdots \\ \beta_0 & \beta_1 x_{1001} & \beta_2 x_{1002} & \beta_1 x_{1003} \end{bmatrix} = \hat{Y}$$

if we have n raws and m columns

$$\hat{Y} = \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_n \end{bmatrix} = \begin{bmatrix} \beta_0 & \beta_1 x_{11} & \beta_2 x_{12} & \cdots & \beta_m x_{1m} \\ \beta_0 & \beta_1 x_{21} & \beta_2 x_{22} & \cdots & \beta_m x_{2m} \\ \vdots & \vdots & \vdots & & \vdots \\ \beta_0 & \beta_1 x_{n1} & \beta_2 x_{n2} & \cdots & \beta_m x_{nm} \end{bmatrix}$$

$$\hat{Y} = \begin{bmatrix} 1 & x_{11} & x_{12} & x_{13} & \cdots & x_{1m} \\ 1 & x_{21} & x_{22} & x_{23} & \cdots & x_{2m} \\ \vdots & \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & x_{n3} & \cdots & x_{nm} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_m \end{bmatrix}$$

$$\hat{Y} = X\beta$$

X = prediction of all input raws matrix

β = all the coefficient matrix

$\hat{Y}$ = prediction matrix

$$Y(\text{actual}) = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \qquad e = Y - \hat{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} - \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_n \end{bmatrix} = \begin{bmatrix} y_1 - \hat{y}_1 \\ y_2 - \hat{y}_2 \\ \vdots \\ y_n - \hat{y}_n \end{bmatrix}$$

in simple linear regression

$$E(\text{error}) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$E = e^T e$$

$$E = \underbrace{\begin{bmatrix} (y_1 - \hat{y}_1) & (y_2 - \hat{y}_2) & \cdots & (y_n - \hat{y}_n) \end{bmatrix}}_{1 \times n} \begin{bmatrix} y_1 - \hat{y}_1 \\ y_2 - \hat{y}_1 \\ \vdots \\ y_n - \hat{y}_1 \end{bmatrix}_{n \times 1}$$

$$E = (Y - \hat{Y})^T (Y - \hat{Y})$$

$$= (Y^T - \hat{Y}^T)(Y - \hat{Y}) \qquad \because (A-B)^T = A^T - B^T$$

$$= \left[Y^T - (X\beta)^T\right](y - X\beta)$$

$$= Y^T Y - Y^T X\beta - (X\beta)^T Y + (X\beta)^T (X\beta)$$

proof $Y^T X\beta = (X\beta)^T Y$

let $Y = A \quad X\beta = B$

$A^T B = B^T A$ ——— (i)

$\because (A^T B)^T = B^T A$ ——— (ii)

$A^T B = (A^T B)^T$

let $A^T B = C$ prove $C = C^T$

$$\text{or } \left(Y^T X\beta\right)^T = Y^T X\beta$$

$$E = Y^T Y - 2Y^T X\beta + \beta^T X^T X\beta$$

$$\frac{dE}{d\beta} = \frac{\partial}{\partial\beta}\left[Y^T Y - 2Y^T X\beta + \beta^T X^T X\beta\right]$$

$$\frac{\partial E}{\partial\beta} = 0 - 2Y^T X + \frac{\partial}{\partial\beta}\left[\beta^T X^T X\beta\right] = 0$$

$$\because \quad y = A^T A$$

$$\frac{dy}{dA} = 2xA^T$$

$$= -2y^T x + 2x^T x\beta^T = 0$$

$$2X^T X\beta^T = 2Y^T X$$

$$\beta^T = Y^T X\left(X^T X\right)^{-1}$$

$$\left(\beta^T\right)^T = \left[Y^T X\left(X^T X\right)^{-1}\right]^T$$

$$\beta = \left[\left(X^T X\right)^{-1}\right]^T \left(Y^T X\right)^T$$

$$\beta = \left[\left(X^T X\right)^{-1}\right]^T \cdot X^T Y$$

$$\boxed{\beta = \left(X^T X\right)^{-1} X^T Y}$$

$\because X$ is square matrix

$X = X\_train$

$Y = Y\_train$

for calculating coefficient and offset we use gradient descent rather than OLS because computational complexity for calculating X^T or Y^T is very high.

in multiple linear regression, the random errors (e) are assumed to be

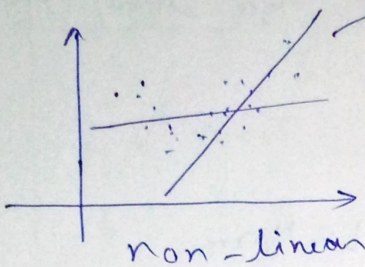
- independently and identically distributed
- normally distributed with zero mean
- constant variance (σ^2)

Date - 17-02-24

polynomial ^ linear Regression.

$y = mx + c$ or $y = \beta_0 + \beta_1 x$ → simple linear Regression.

$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots \beta_n X_n$ → multiple linear Regression



non-linear data.

Here linear Regression also work but result are not good. Hence here we use polynomial linear Regression.

Hence for solve this we add polynomial term in the given data

example

X	y
35	100

→

X^0	X^1	X^2	Y
1	35	1225	100

degree = 2 → hyperparameter

For polynomial Regression

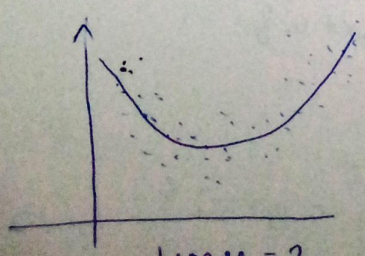
for degree = 2 : $y = \beta_0 + \beta_1 X + \beta_2 X^2$

if colum | X_1 | X_2 | Y. | : $y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \beta_3 X_2 + \beta_4 X_2^2$

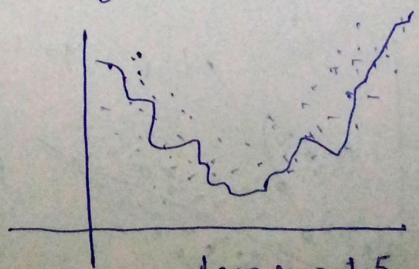
Here $y \propto \beta_0, \beta_1, \beta_2, \beta_3, \beta_4$

linearly dependent

Here we called it as polynomial linear Regression



degree = 2



degree = 15