Linear Regression

A very easy to understand algorithm, supervised machine learning algorithm.
we mostly work on where O/P variable is "numerical"

3 Types
1.) Simple Linear Regression
2.) multiple " "
3.) Polynomial " "

Simple Linear Regression

1 input column 1 O/P column

ex

CGPA	Package

→ For some CGPA we are trying to predict what job compensation ("package") a person would get.

Multiple Linear Regression

we have more than 1 input variable
for ex CGPA | STATE | GENDER | IQ | PACKAGE
if we use starting 4 columns to predict package

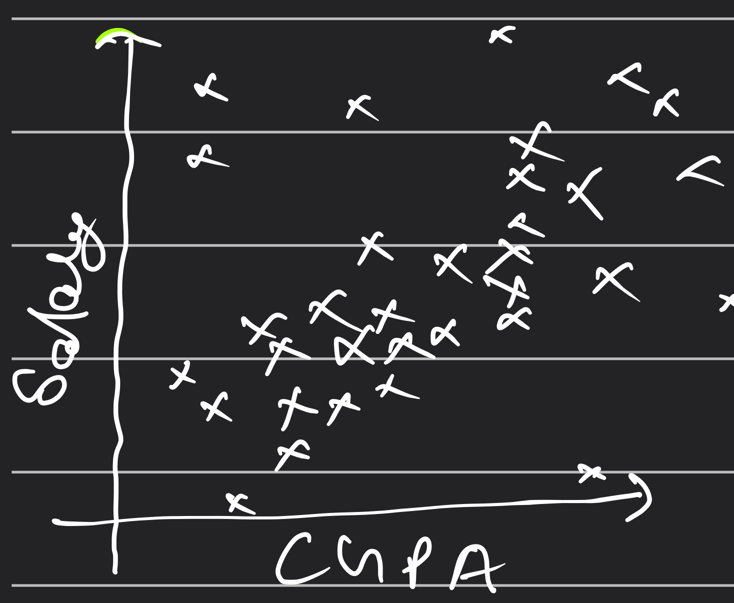
Polynomial

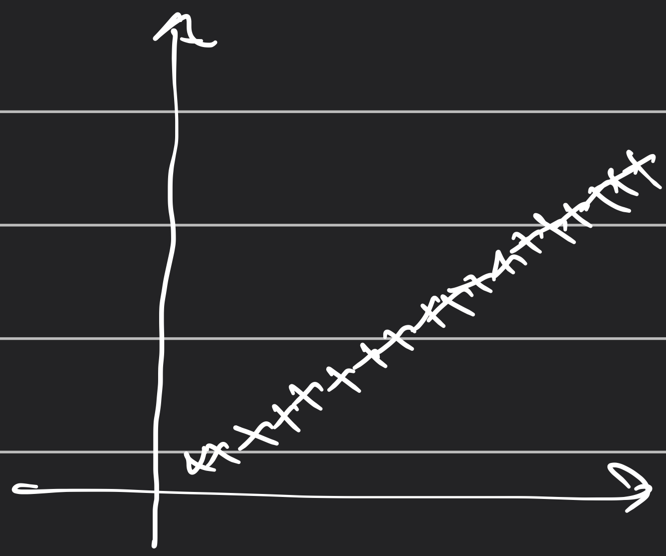
for now just remember, when we dont have data i.e not linear in nature we use polynomial.

Δ what if I didn't had the data & someone asked me, what package/salary people get in your college?

⇒ One answer would be give the Average of all packages.

Δ Sort of Linear vs Completly Linear



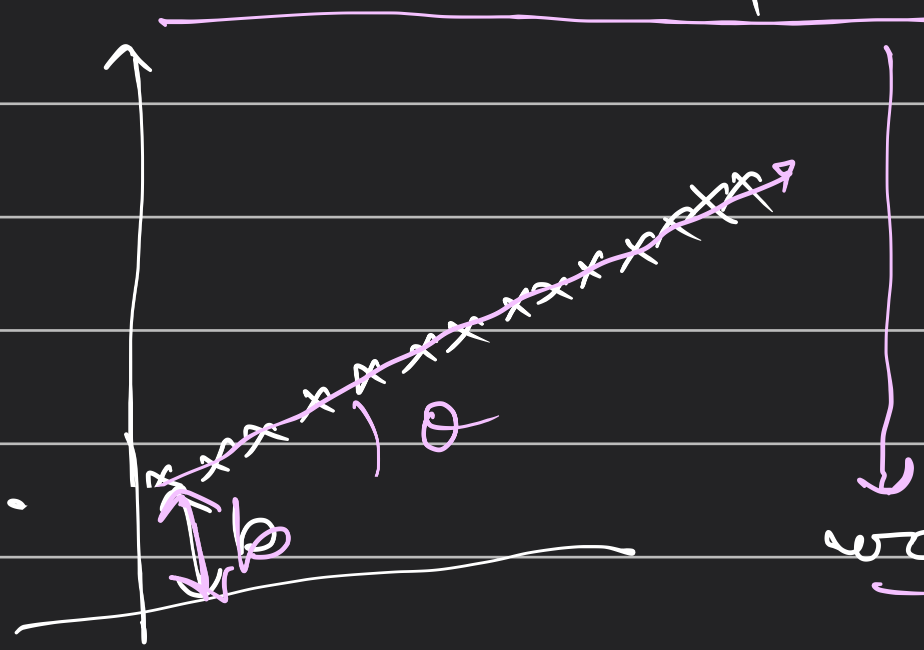


• Real world data getting completly linear is rare.

force → Some got amazing GPA but less salary and vice versa. This stohastic error ⇒ Errors that we cannot calculate.

Real world data gets affected by real world errors which cannot be quantified.

If we had completly Linear data



Eq Line ⇒ $y = mx + b$

b = y intercept

for each "x" i will get a "y". This would have so easy

But what do in case of Sort of Linear data

⇒ Perfect Line → a line that will go through each line → feasable for completly linear data but for us we use "Best fit Line" → this line is doing the less number of errors, a line that goes very close to each point. i.e find slope (m) and intercept (b) such that line goes closes to each point.

$\boxed{\text{Conclusion → In linear regression we try to draw a line that is close passing through each point.}}$

- Intuition

$$y = mx + b$$

package = m * cgpa +

"m" → weightage → the dependence of input variable on slope

If m is less than CGPA wont have much impact i.e package/salary is not much dependent on CGPA.

If slope is more than some variation on change in CGPA will have im impact on salary.

"b" ⇒ offset, even if input to variable cause no effect our output will still has something to give.

For salary = m * Experience + b   if b=0

And for Fresher experience is 0 $\begin{array}{c|c} exp & salary \\ 0 & \end{array}$

that means salary = 0 but this doesn't happen so "b" i.e offset creates this difference that even a fresher would have some salary. For example you can call this as "Minimum Wage". even if you have experience or not you will get your minimum wage.

$ How do we calculate the value of "m" & "b"?

Closed Form Solution
↳ we use direct formula
Ordinary Least Square technique

Non closed form solution ⇒ We use approximation techniques
↳ Gradient Descent.

why use approximation technique like gradient descent when we have direct formula?

Internally Scikit learn uses OLS (ordinary least square technique), but when we have higher dimension data we use gradient descent. "SGD Regressor" class is used to utilize gradient descent.

- OLS

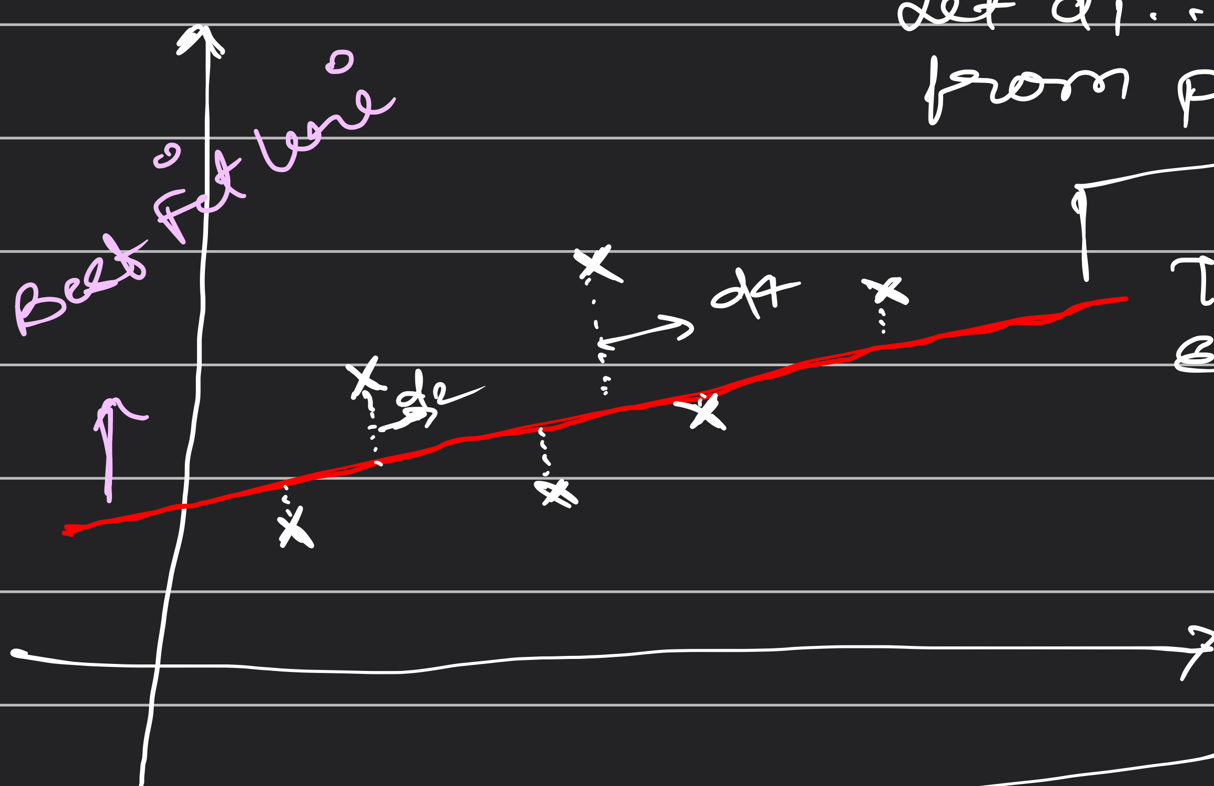
Calculate

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

$\bar{x} \rightarrow$ mean

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

$\bar{y} \rightarrow$ mean



Let $d_1 \ldots d_6$ be distance from point to line

Total Error $E = d_1 + d_2 + d_3 \ldots\ldots + d_n$

$E = d_1^2 + d_2^2 + \ldots d_n^2$

$\rightarrow$ so that all distances are +ve

why not use Mod $\Leftarrow$

$\Rightarrow$ To penalize outliers

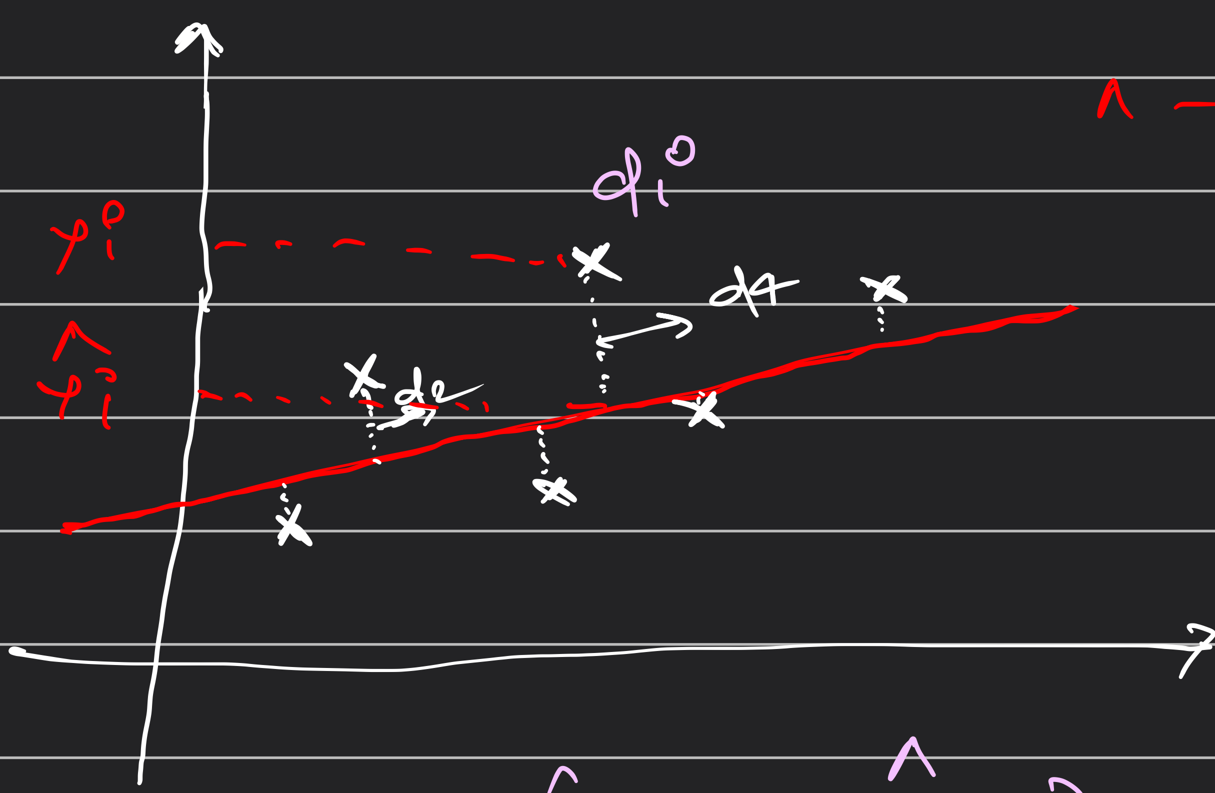
$\Rightarrow$ Differentiation at origin is not possible

$$\boxed{E = \sum_{i=1}^{n} d_i^2} \rightarrow \text{Error function}$$

J in some book represented

$\Rightarrow$ So our goal is to search of (m, b) that minimizes E

So now



^ → predicted value

$d_i = (y_i - \hat{y}_i)$ ⇒ Error for a particular point in terms of y value

So Error

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

Average $= \frac{E}{n}$ = Average Error.

n → no of observatn.

In terms of (m, b) minimized $\hat{y}_i = m x_i + b$

$$E_{(m,b)} = \sum_{i=1}^{n} (y_i - m x_i - b)^2$$

So we to minimize we found value m and b.

Here x_i and y_i are fixed → actual package*

CGPA*

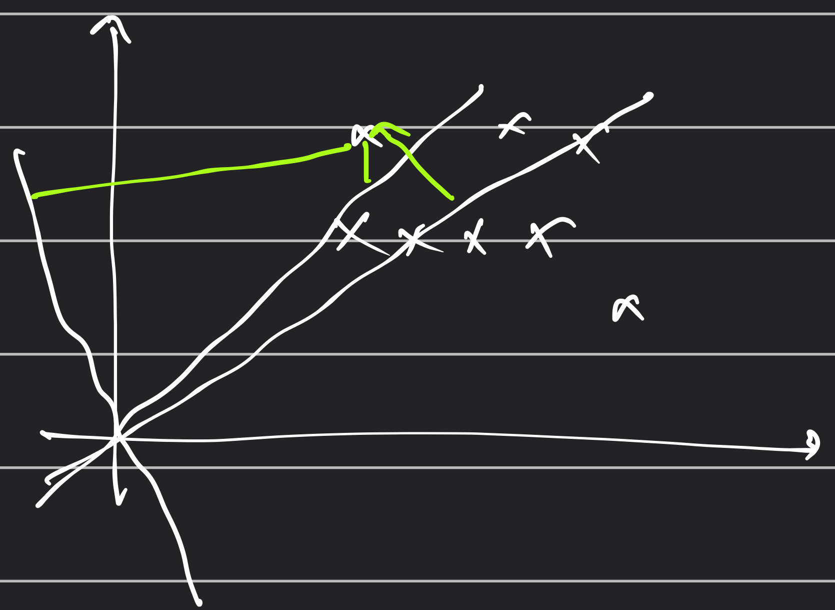
From this equation we understood that any change m or b or m and b will have effect on E

⇒ Lets see how "m" effects E

Suppose $b = 0$

$$E_{(m)} = \sum_{i=1}^{n} (y_i - m x_i)^2$$

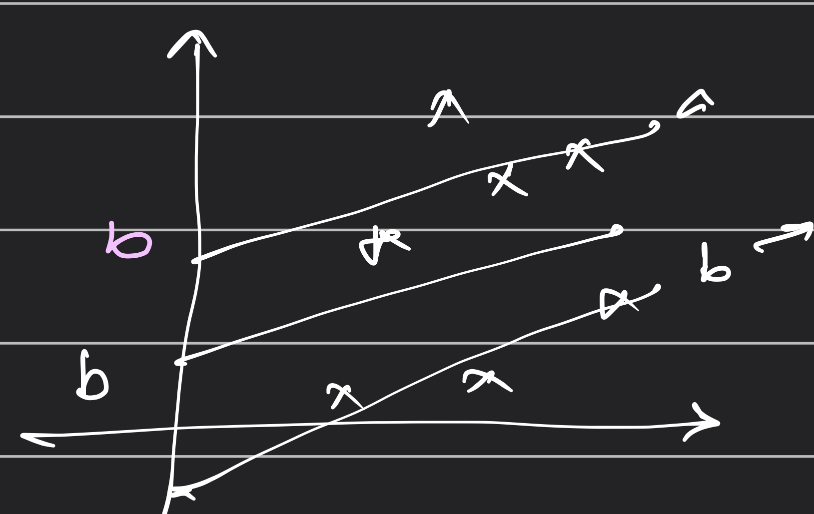
now as "b" i.e intercept on y is 0 our line will go through origin.

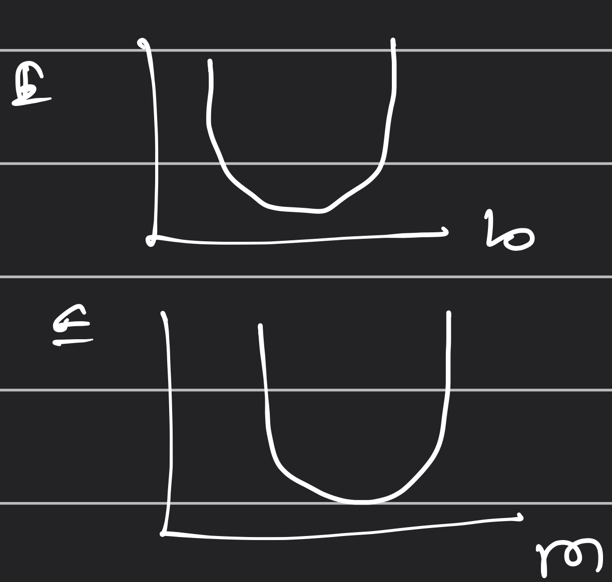


So m gets rotated accordingly to minimize b

⇒ say m is constant = 1

$$E_{(b)} = \sum_{i=1}^{n} (y_i - x_i + b)$$





For certain value m & b E is max & certain value E is less but How to find those "m" & "b"

⇒ This is where calculus comes into play i.e Maxima & Minima derivates.

$$\frac{dE}{dm} = 0$$

To find minimized m ~~b~~

Derivate of Error w/ respect to slope m

To find minimized "b"

Derivate of ~~intercept~~ Error w respect to ~~error~~ intercept

$$\frac{dE}{db} = 0$$

→ we find minimum value at 0

$$\frac{dE}{db} = \frac{\partial}{\partial b} \sum_{i=1}^{n} (y_i - mx_i - b)^2 = 0$$

$$2\sum 2x\left(\frac{\partial y_i}{\partial b} - \frac{\partial mx_i}{\partial b} - \frac{\partial b}{\partial b}\right) = 0$$

$$\times (y_i - mx_i - b)$$

$$= \sum 2 (0 - 0 - 1)(y_i - mx_i - b)$$

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

$$= \sum^{n} (y_i - mx_i - b = 0$$

$$= \sum_{i=1}^{n} y_i - \sum mx_i - \sum b = 0$$

Divide by total n → observation

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

$$\Rightarrow \bar{y} - m\bar{x} - \frac{nb}{n} = 0$$ this how we got this eq

$$\Rightarrow \bar{y} - m\bar{x} = \frac{\cancel{n}b}{\cancel{n}} \Rightarrow b = \bar{y} - m\bar{x}$$

Lets see how we get eq of m

$$E = \sum (y_i - mx_i - b)^2$$

$$E = \sum \left(y_i - mx_i - (\bar{y} - m\bar{x})\right)^2$$

$$\frac{dE}{dm} = 0$$

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

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

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

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

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