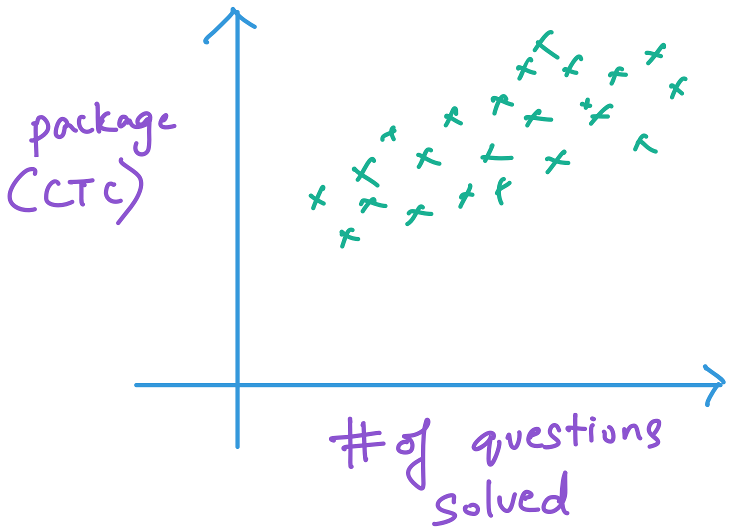
Regression

$$Y = f(X)$$

Dependent variable
(Number)
Continuous

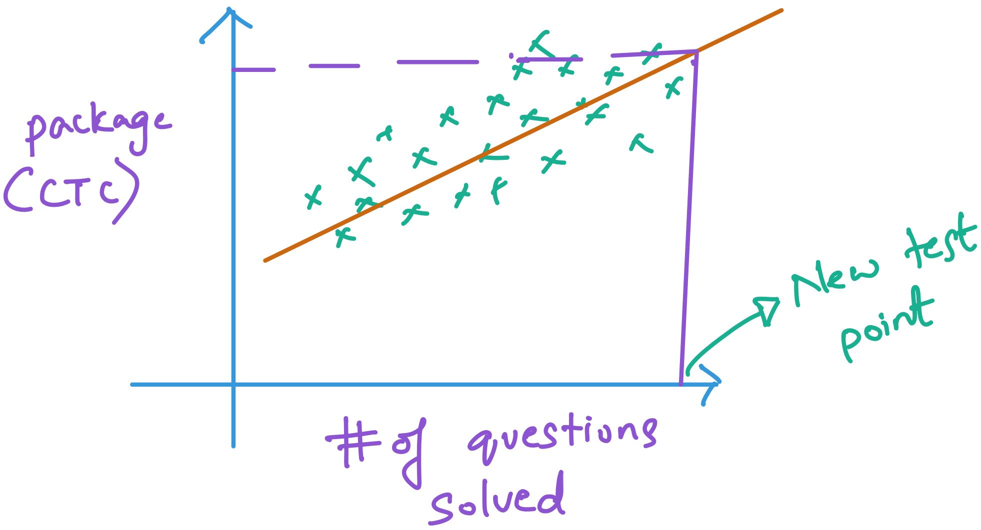
Independent variable
(features)



Can't draw 1 line to fit all points

⇒ "Best fit" line

"Linear Regressor"



$$y = wx + b$$

co-efficient

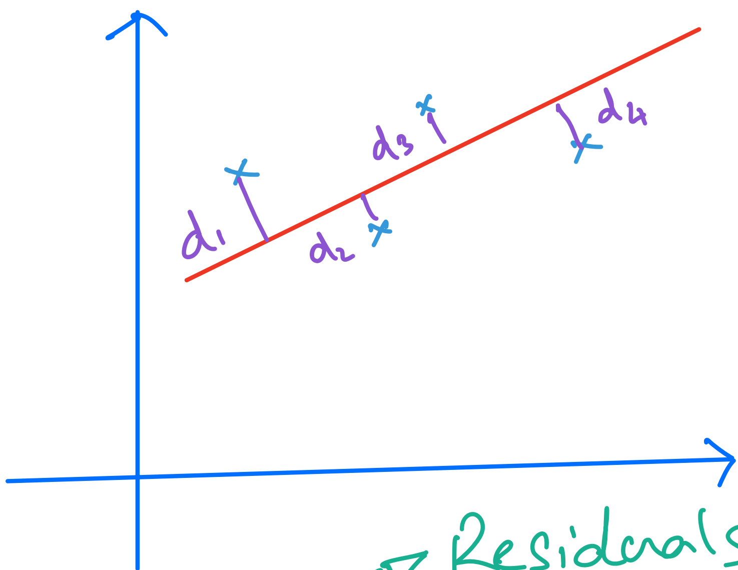
intercept (bias : o/p in the absence of input)

x(# of q solved)	y(CTC)
20	6
35	9
50	14
70	19
100	?

How do you solve for w & b ?

① Ordinary Least Squares

– Closed form solution



If $\left.d_1 \cdots d_4\right\} = 0$

$\Rightarrow$ Perfect fit!

Find a line that minimizes this!

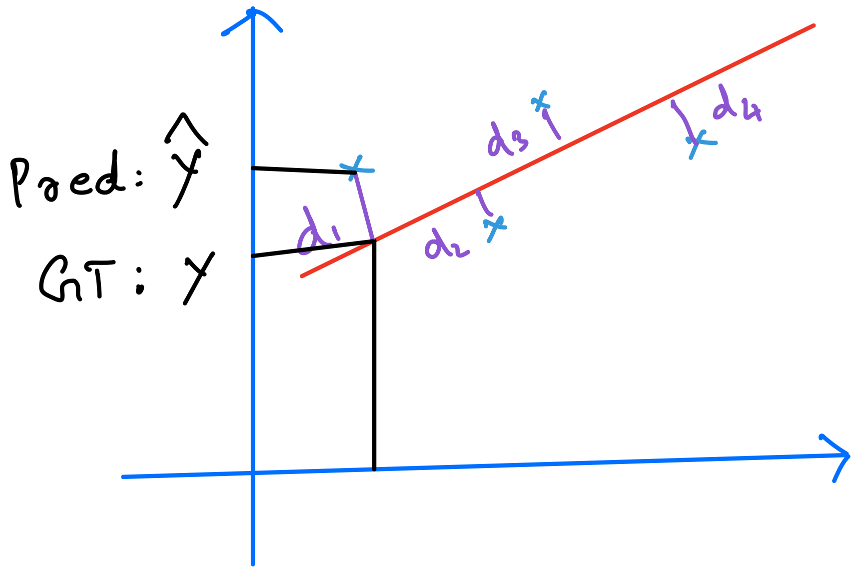
Residuals

$$J = d_1^2 + d_2^2 + \cdots + d_n^2$$

$$J = \sum_{i=1}^{n} d_i^2$$

Find 'm' & 'b' that min. J

where are m & b?



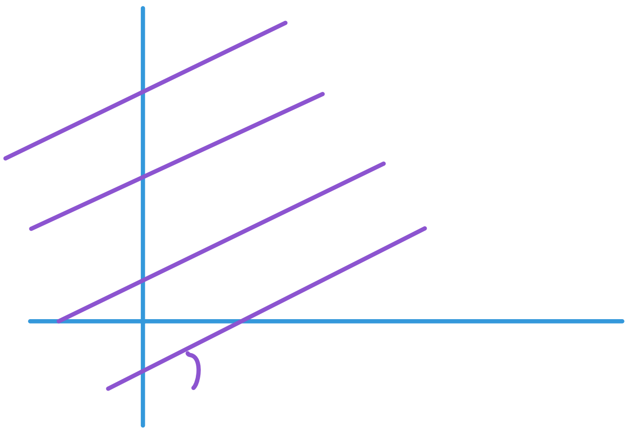
$$d_i^2 = \underbrace{(x_i - \hat{x}_i)^2}_{0} + (y_i - \hat{y}_i)^2$$

$$\Rightarrow d_i^2 = (y_i - \hat{y}_i)^2$$

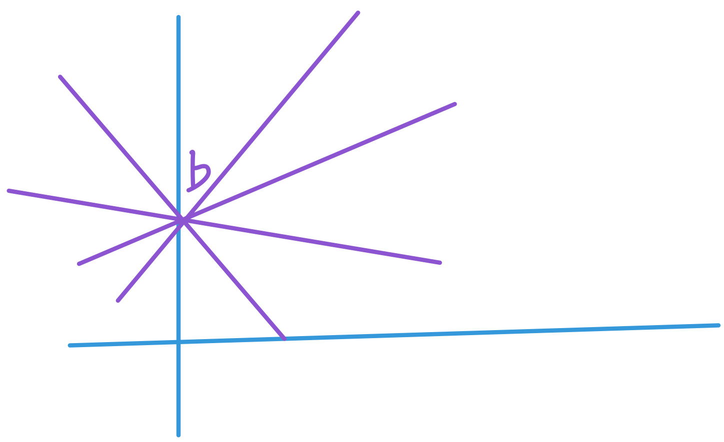
$$\therefore J = \sum_{i=1}^{n} (y_i - \underbrace{\hat{y}_i}_{\to\, wX_i + b})^2$$

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

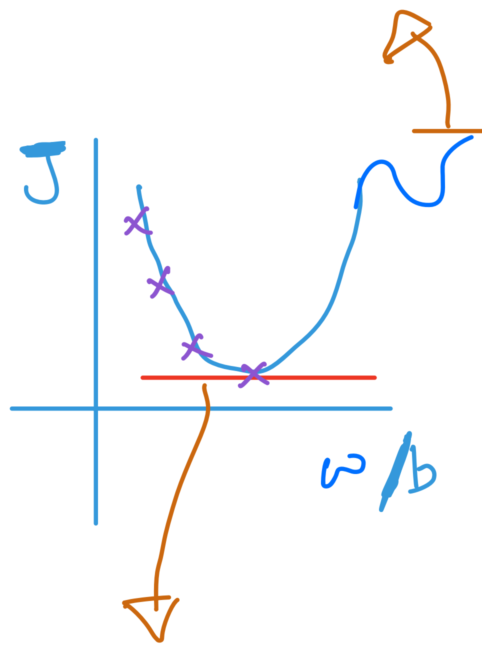
① If w is constant :



② If b is constant



Maxima



Minima
(slope = 0)

$$\frac{\partial J}{\partial w} = 0 \qquad \frac{\partial J}{\partial b} = 0$$

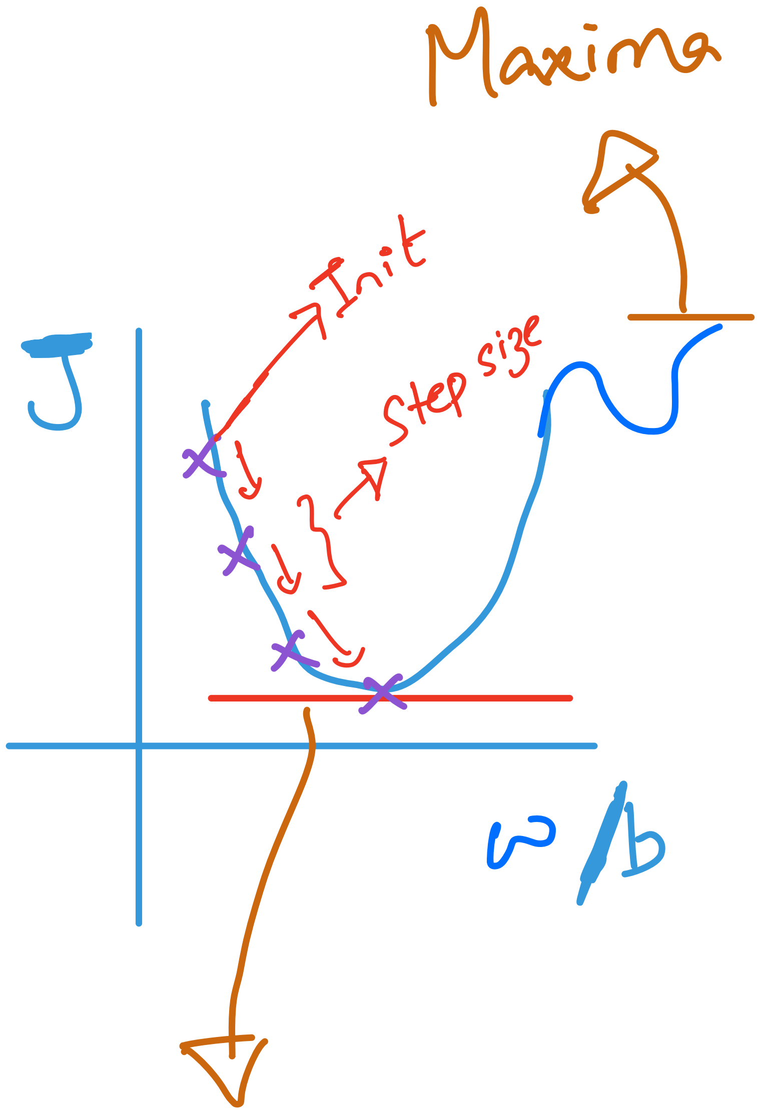
$$w = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

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

$\bar{x}, \bar{y}$: Mean.

② Gradient Descent

— Not closed form solution



- Higher dimensional space
- Approximation algo

Multiple Linear Regression

X : # of questions solved
College
Cgpa
Company

Y : CTC

ordinal data

$$Y = \underbrace{w_1 X_1}_{\#q} + \underbrace{w_2 X_2}_{\substack{small=0 \\ med=1 \\ large=2}} + \underbrace{w_3 X_3}_{\#\ cgpa}$$

$$+ w_4 X_4$$
$$+ w_5 X_5$$
$$+ w_6 X_6$$
$$+ w_7$$

Company
one-hot
001 → google
010 → MS
100 → apple

categorical

Collinearity

How features are related?

\# questions solved $\propto$ college

"Variance Inflation Factor"

Checks correlation

1 : no collinearity

1-5 : moderate

≥ 5 : severe

Mitigation strategy

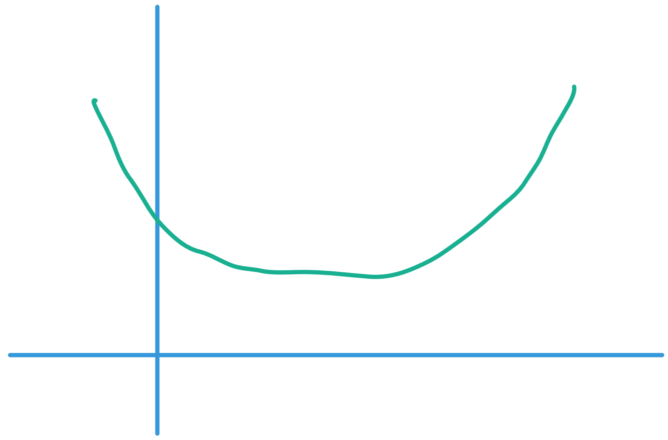
① Centering features

$X = X - \bar{X}$

Lowers VIF

$$Y = w_1 x_1 + w_2 x_2 + w_3 x_1 x_2 + w_4 x_1 x_1$$

Correlation



Multiple : Matrices

X_s Calc derivatives to minimize J

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

$$MSE = \frac{1}{m} \sum_i (\hat{y}_i - y_i)^2$$

Mean Squared Error

$$= \frac{1}{m} \|\hat{y} - y\|_2^2$$

Minimize MSE i.e, gradient/slope = 0

$$\nabla MSE = 0$$

$$\Rightarrow \nabla \frac{1}{m} \|\hat{y} - y\|_2^2 = 0$$

$$\Rightarrow \frac{1}{m} \nabla_w \|Xw - y\|_2^2 = 0$$

$$\Rightarrow \frac{1}{m} \nabla_w (Xw - y)^T (Xw - y) = 0$$

$$(x^2 = x \cdot x = x^T x)$$

$$\Rightarrow \boxed{w = (X^T X)^{-1} X^T y}$$

Normal Equations

Types of Regression

① Polynomial regression

$$y = w_1 x_1 + \underbrace{w_2 x_2^2}_{\text{Quadratic in } X_2} + w_3 x_1 x_2$$

BUT

Linear in w !!!

$\therefore$ Normal Equations can be used

② Non-linear Regression (in w)

$$y = w_1 x_1 + w_2^2 x_2 + w_1 w_2 x_3$$

$$y = \log(w_0 + w_1 x)$$

How has the model improved?

① More features

$\Rightarrow$ More parameters ('w')

② Representational Capacity

(i) Polynomial in X

(ii) Polynomial in W