Simple Linear Regression

x	y

↳ y is continuous.

Relationship Between x & y is linear.

Hyper params

— We have to decide before building the model

$$y = \beta_0 + \beta_1 x$$

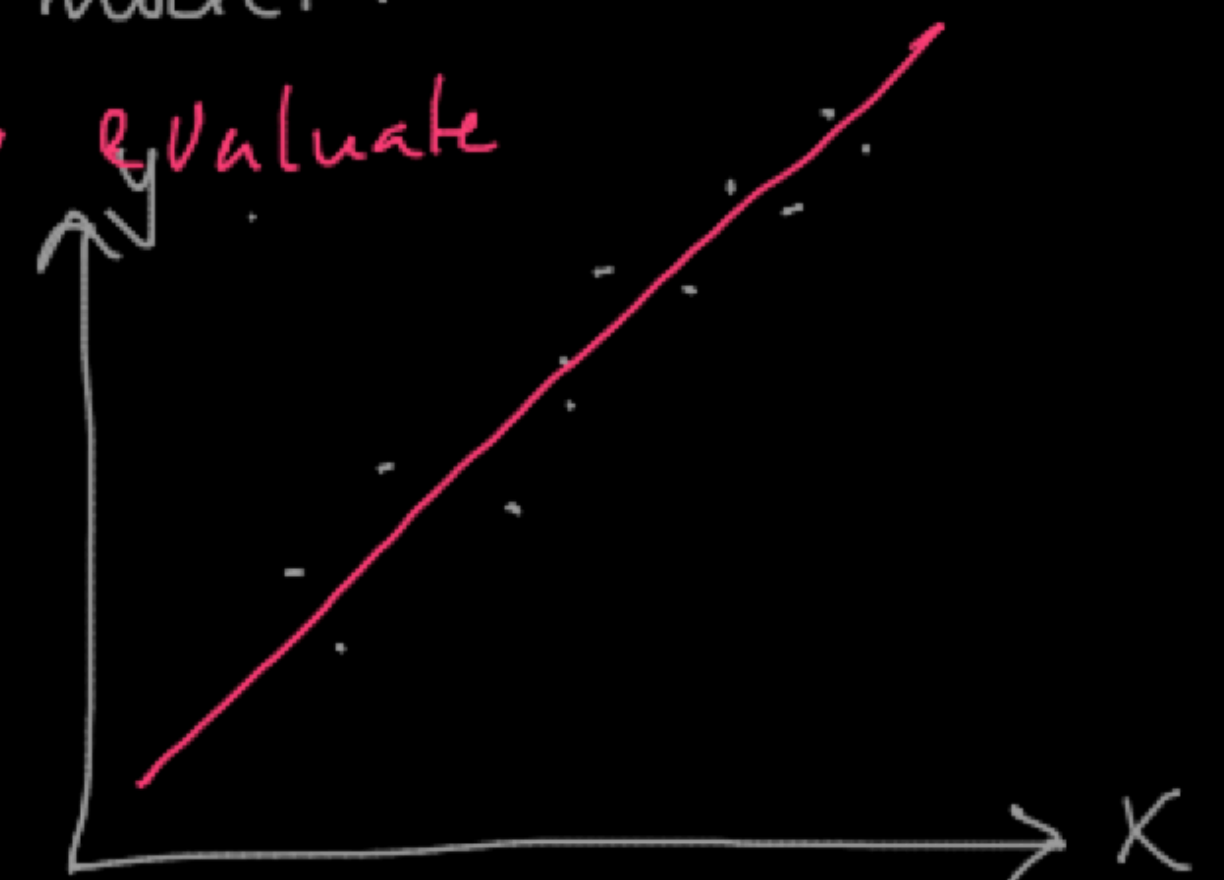
$\beta_0, \beta_1 \rightarrow$ Model parameters.

→ They are estimated by the model.

Task of the algorithm: (OLS)

Ordinary Least Squares.

1. Visually evaluate



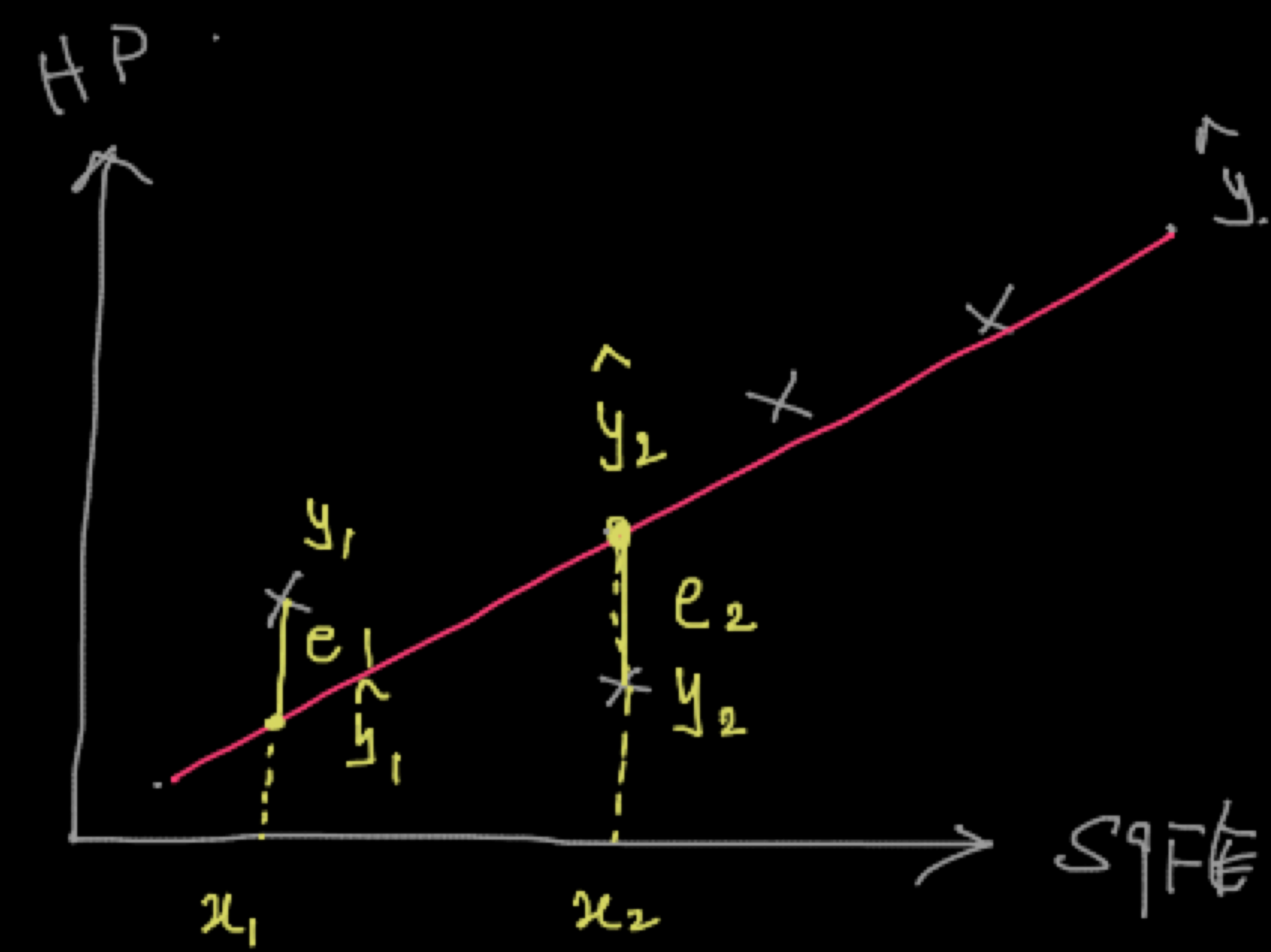
2. Quantify the strength of the relationship.

$r \Rightarrow$ Pearson's Correlation Co-efficient.

$r \geq \pm 0.85$

→ represents the strength of the LINEAR relationship betw x & y.

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



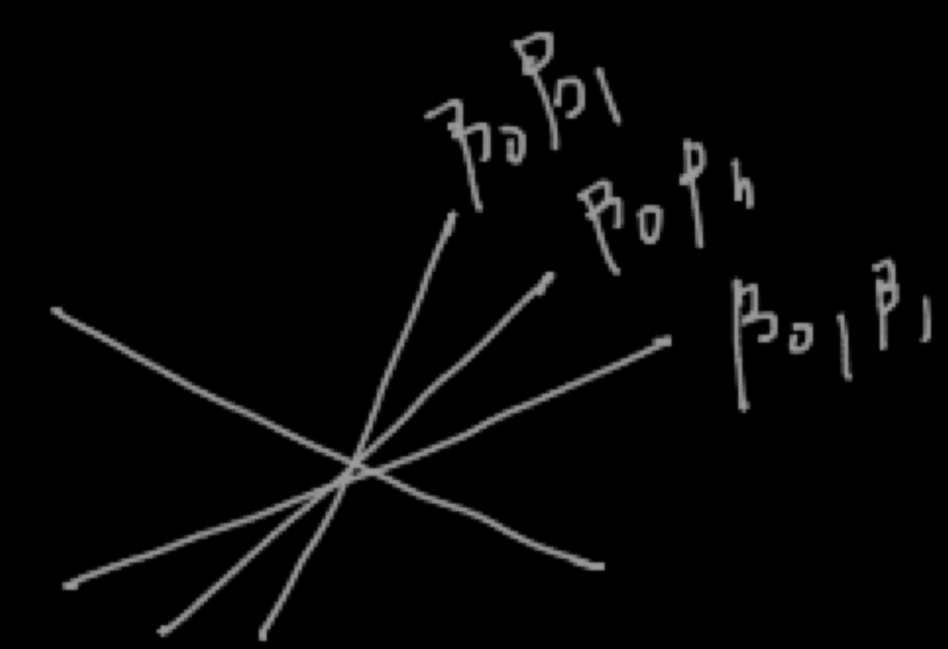
$$\hat{y} = \beta_0 + \beta_1 x$$

$y - \hat{y} \rightarrow$ Residual (Error).

Loss fn $\rightarrow$ Some fn of the Error or residual.

x	y	$\hat{y}$	$\Sigma(y-\hat{y})^2$
x_1	y_1	$\hat{y}_1$	$(y_1 - \hat{y}_1)^2$
x_2	y_2	$\hat{y}_2$	$(y_2 - \hat{y}_2)^2$
x_3	y_3	$\hat{y}_3$	$(y_3 - \hat{y}_3)^2$
x_4	y_4	$\hat{y}_4$	.
x_n	y_n	Σ	SSE

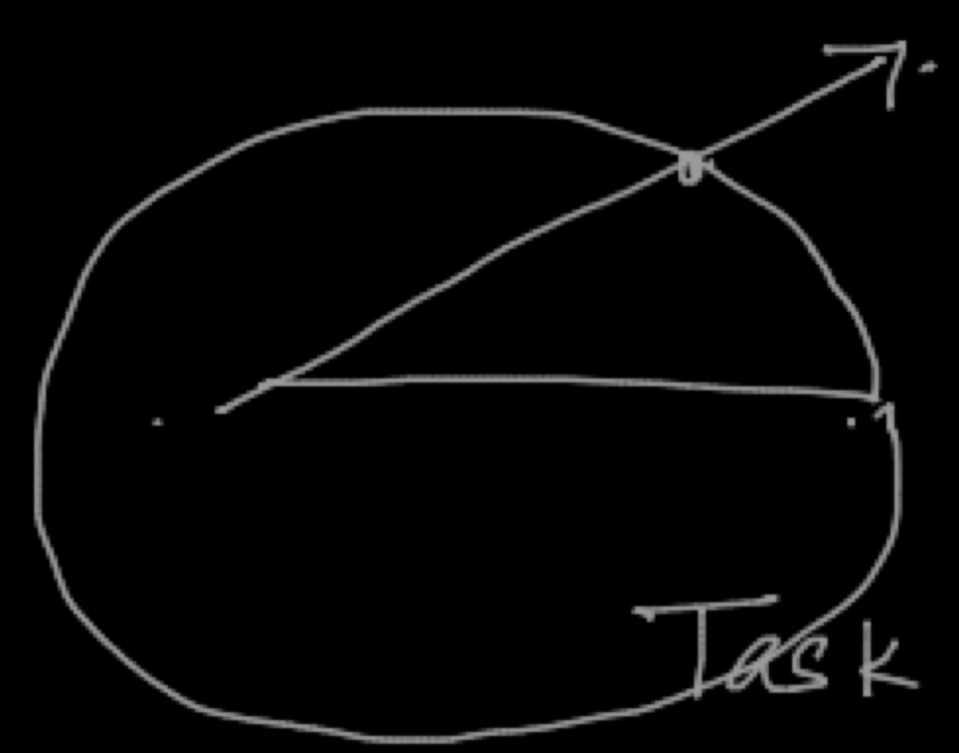
$$SSE = e_1^2 + e_2^2 + e_3^2 + \cdots e_n^2$$



$$\downarrow (SSE) = \sum_{i=1}^{n} \left[y_i - [\beta_0 + \beta_1 x_i] \right]^2$$

(L)

$\Rightarrow$ loss fn (squared loss)



Task of the algorithm:

'Those β's which will minimize the squared loss' $\rightarrow$ Gradient Descent

$\rightarrow$ OLS.

How well is the model doing.

1. R^2 ← What is the percentage of Variation explained by the features in the model.

2. Metrics — MSE.
→ RMSE.
→ MAE.
→ MAPE

$\boxed{R^2 \rightarrow 85\%}$ → 90% → 98%. 1L

75L 77 78

20000

7.78L

Feature

#Bed	SqFt x_1	HP (y)
	800	30
	100	47
	700	23
	2500	65
	3000	97
	3000 1000 →	100

$\bar{y} =$

$\bar{y} =$ 50 Lakh

Unexplained Variation → 1L [78L → 2500 } , 75L , 2500.

$\hat{y}$

$\bar{y}$ → 50 L 75L

Explained Variation

50 L 35L → 1500

Variation in price

$$R^2 = \frac{(SSR)}{(SSR+SSE)} \rightarrow \frac{\text{Explained Variation } (\hat{y}-\bar{y})}{\text{total Variation}}$$

(Expl + Unexpl Variation)

$(y-\hat{y})$

ε

Sqft
x y

$\bar{y} \Rightarrow$ Avg Value

absolute

x	y	$\hat{y}$	$(y-\hat{y})$ Residual	$(y-\hat{y})^2$	$\lvert(y-\hat{y})\rvert$
x_1	y_1	$\hat{y}_1$	$(y_1-\hat{y}_1)$	$(y_1-\hat{y}_1)^2$	+ve $\lvert(y_1-\hat{y}_1)\rvert/y_1$
x_2	y_2		+ve		+ve $\lvert(y_2-\hat{y}_2)\rvert/y_2$
x_3	y_3		−ve		+ve
x_n	y_n			$(y_n-\hat{y}_n)^2$	

$\Sigma \rightarrow$ SSE

$\frac{\Sigma}{n}$ = MAE MAPE

✓ MSE $= \frac{SSE}{n} \Rightarrow$ 4(Lakh^2)

✓ RMSE $= \sqrt{MSE} \rightarrow$ (2 lakh) $\pm$ 2L

=

Machine Learning (supervised)

Regression — Classification

→ y is Continuous — y is categorical

x	y	$\hat{y}$	$E = (y - \hat{y})$		$(y-\hat{y})^2$	$\lvert y - \hat{y} \rvert$
x_1	y_1	$\hat{y}_1$	$(y_1 - \hat{y}_1) = +ve$	$+2L$	4	2
x_2	y_2	$\hat{y}_2$	$(y_2 - \hat{y}_2) = -ve$	$-1L$	1	1
x_3	y_3	$\hat{y}_3$	$(y_3 - \hat{y}_3) = -ve$	$-2L$	4	2
⋮	⋮	⋮	⋮		⋮	
x_n	y_n	$\hat{y}_n$	$(y_n - \hat{y}_n) = +ve$	$+1L$	1	1
				$\times 0$	$10L = SSE$	$6/4 = MAE$

$MSE = \left(\frac{10}{4}\right) = (2.5)L^2$

$RMSE = \sqrt{MSE} = 1.58L$ — $35 = \begin{cases} 36.58 \\ 33.42 \end{cases}$

1. Defined a loss fn. ✓

SSE ⇒ Squared loss.

2. We will find those model params which will minimize the loss. ✓

Break

$\boxed{10.10 \text{ am}}$

← X →

Cholestr x_2	Waist circum (x)	Adipose tissue (y)
—	—	—

. csv.

$\beta_1 = 0$.

$y = \beta_0 + \beta_1 x$ ↑ 1 unit.

in 'y'.

$$y = \boxed{\beta_1 x_1 + \beta_2 x_2} + \beta_0$$

x_1	x_2	x_3	y
1	1	1	y_1
2	1	1	y_2

$\}\ \beta_1$

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

0

$y = f(x)$

↳ linear eqn

$\beta_2 = 0$ $\beta_1 = 0$

$H_0: \beta_1 = 0;\ H_a: \beta_1 \neq 0.$

$p = 0.0 \rightarrow x_1$ is a good predictor.

$H_0: \beta_2 = 0;\ H_a: \beta_2 \neq 0.$

$p = 0.43 > 0.05 \rightarrow \beta_2 = 0;$

x_2 is Not a good pred