Linear Regression

18 May 2025 07:20

Linear Regression

- Supervised ML technique
- Solves the Regression Problems
- Models the relationship b/w independent variables (features/input variables) and a dependent variable (i.e. target/outcome) by fitting a linear equation to the observed data.

Goal :

New Input (Features) → ML → Outcome (Target) → Mathematical Representation

Regression Model

Imp

Model a relationship
⇓
Tries to understand a pattern out of the data.

Regression Problem

Regression problems aim to predict continuous, numerical values like price or weight, based on one or more input features as opposed to discrete categories like in classification problems.

https://www.kaggle.com/datasets/yasserh/housing-prices-dataset

Examples :
- Predicting house prices based on features like size, location, and number of bedrooms.
- Predicting the temperature based on weather data.
- Predicting stock prices based on various market factors.

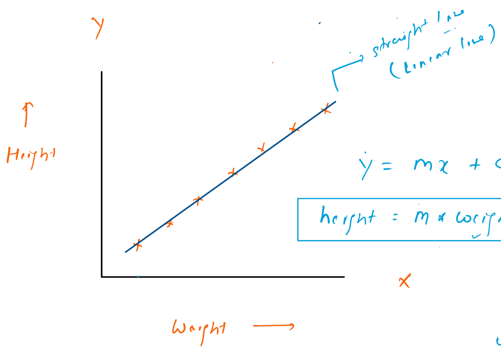
Sales of a car → prediction → how many units will be sold in 2025-26

House price → what is the price of a house in a given locality

Not a Regression problem

8 Email. → Spam (1) / Ham (0) — categorical

I/p = Flower → o/p Setso / Versicolor / Virginica — categorical



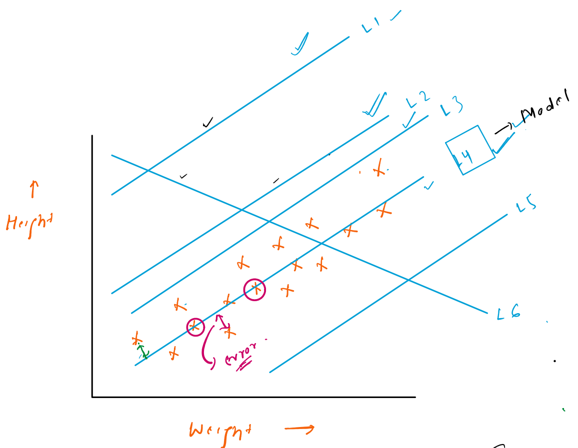
Height   Weight

$y = mx + c$

$\boxed{height = m * weight + c}$ → Mathematically.

$y = b_1 x + b_0$

y → Dependent; b_1 → slope; x → independent; b_0 → intercept.



① Line 4 passes closely to all the points. [Best line the approximates the data] → Best fit Line.

① Line 4 explains the points in a better manner.

Mathematically.

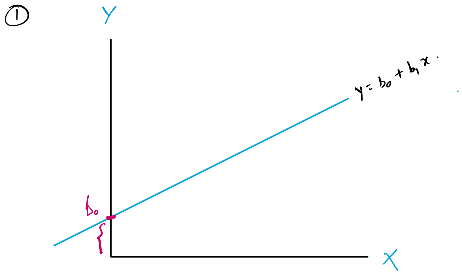
$\boxed{height = b_1 * weight + b_0}$ → Model

To explain the behaviour —
① b_0
② b_1

To explain the behaviour –

1. b_0
2. b_1
3. How to get the best fit line ?

1



$$\boxed{y = \underbrace{b_0}_{\text{intercept (y}} + \underbrace{b_1}_{\text{slope}} * x.}$$

– Even if, the value of x is zero y will have some value.

$salary = b_0 + b_1 * experience$

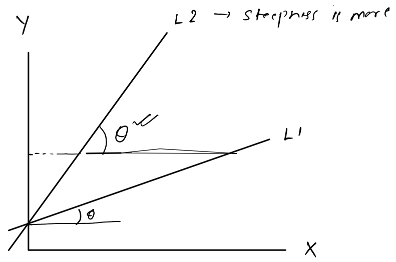
$\Rightarrow salary = b_0 + b_1 * 0$

$\Rightarrow \boxed{salary = b_0}$

$\underline{\underline{b_0 = 20k, 30k, 40}}$

2. Slope (b_1) => coefficient of the equation.

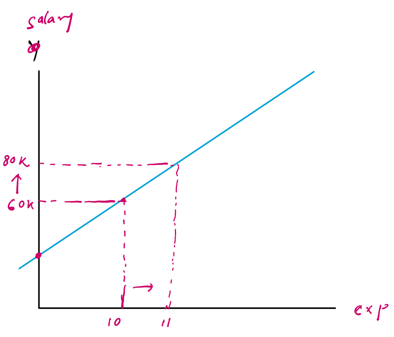
$y = b_0 + \underbrace{b_1}_{\text{slope.}} * x$



$salary = b_0 + \boxed{b_1} * exp$

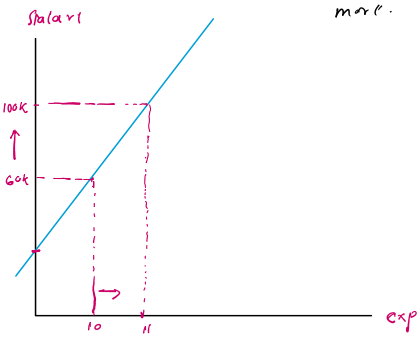
$30 + 2 * 1 = 32\ K$

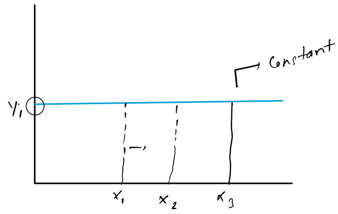
$30 + \underline{\underline{10}} * 1 = \underline{\underline{40\ k}}.$



1. change in y for a unit change in x

2. when the slope is less, change will be less. And when the slope is more, change will be more.





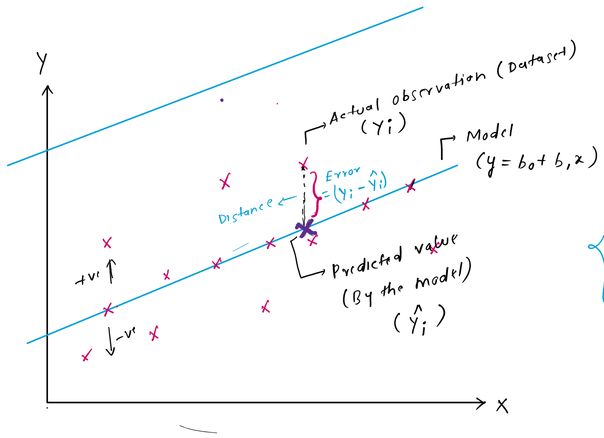
③ Techniques to find best fit line

① OLS → Ordinary Least Squares.
→ Dataset is small.

② Gradient Descent.
→ Large dataset.

Designed to work using "Differentiation"

Ordinary Least Squares



Actual observation ⇒ Y_i
Predicted value ⇒ $\hat{Y}_i$

① Figure out the errors. $(Y_i - \hat{Y}_i)$

② Since, +ve & -ve distances might cancel out each other so, we take the squares of the differences & sum them.

③ Pick the minimum value.

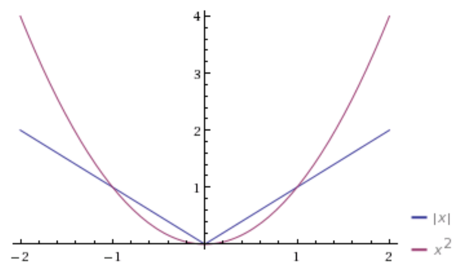
Squares

$$\text{Line 1} \Rightarrow (-1)^2 + (+1)^2 + (-2)^2 + (+2)^2$$
$$1 + 1 + 4 + 4 = 10 \checkmark$$

$$\text{Line 2} \Rightarrow (-1)^2 + (+2)^2 + (-2)^2 + (+3)^2$$
$$= 1 + 4 + 4 + 9 = 18 \times$$

Mod

|-1|	|+1|	|-2|	|+2|	
1	1	2	2	= 6



$$\text{MIN}\left[\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2\right] \Rightarrow$$

Minimum sum of squared distances between the OBSERVED value & the PREDICTED value.

⇓

Many lines

⇓

Pick the one which makes minimum error.

X (~~Exp.~~)	Y (~~Salary~~)
1.1,	39343.00
1.3,	46205.00
1.5,	37731.00
2.0,	43525.00
2.2,	39891.00
2.9,	56642.00
3.0,	60150.00
3.2,	54445.00
3.2,	64445.00
3.7,	57189.00
3.9,	63218.00
4.0,	55794.00
4.0,	56957.00
4.1,	57081.00
4.5,	61111.00
4.9,	67938.00
5.1,	66029.00
5.3,	83088.00
5.9,	81363.00
6.0,	93940.00
6.8,	91738.00
7.1,	98273.00
7.9,	101302.00
8.2,	113812.00
8.7,	109431.00
9.0,	105582.00
9.5,	116969.00
9.6,	112635.00
10.3,	122391.00
10.5,	121872.00

Y → y_train.

X_train, y_train → Train (for model training) ✓

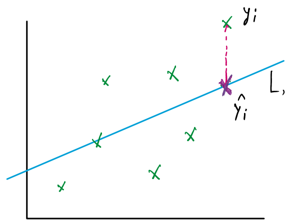
X_test, y_test → Test (for validating the model)

$30 \times 0.2 = 6$ data points for test

Evaluating the model's performance

① R^2 (R-squared) => Coefficient of Determination

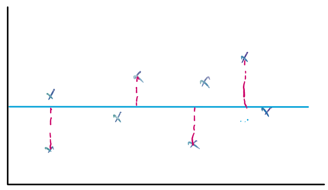
$y_i - \hat{y}_i$ => error
=> distance
=> residual



$\text{sum}(y_i - \hat{y}_i)^2$ => min

$SS_{res} = \text{SUM}(y_i - \hat{y}_i)^2$ —— ①

SS_{res} → sum of squared residuals.



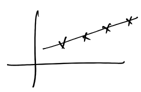
$SS_{TOT} = \text{SUM}(Y_i - Y_{avg})^2$ —— ②

SS_{TOT} → Total sum of squares

R-squared or Coefficient of Determination =>

$$\boxed{R^2 = 1 - \frac{SS_{res}}{SS_{tot}}}$$

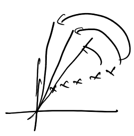
SS_{res} → We always try to reduce it



Case 1

$SS_{res} = 0$ => ~~the~~ Line passes through all the points

$R^2 = 1 - \frac{0}{SS_{total}} = 1$ $\quad \boxed{R^2 = 1}$ ✓



Case 2

SS_{res} increases => R^2 will decrease
=> Going away from 1

$= 1 - \frac{49}{100} = 1 - 0.4$
$= \boxed{0.6}$ ✓

Case 2

SS_{res} increases $\Rightarrow$ R^2 will decrease

$\Rightarrow$ Going away from 1

$\Rightarrow$ The model is bad.

$1 - \frac{40}{100} = 1 - 0.4 = \boxed{0.6}$

$1 - \frac{60}{100} = 1 - 0.6 = 0.4$

$1 - \frac{80}{100} = 1 - 0.8 = 0.2$

Becoming Better $\longrightarrow$

$-ve$ 0 $\boxed{R^2 = 1}$

$\longleftarrow$ Becoming worst

$y = b_0 + b_1 x \Rightarrow R^2$. (Simple Linear Regression)   — SS_{res}.

$y = b_0 + b_1 x_1 + \boxed{b_2 x_2} \ldots\ldots + b_n x_n$ (Multiple Linear Regression)   — color; SS_{res}.

$\Downarrow$

R^2

① $R^2 \Rightarrow$ increases / stays same.

Adjusted R^2 $\Rightarrow$ Modified R^2.

$$\boxed{Adj\ R^2 = 1 - (1 - R^2)\frac{n-1}{n-p-1}}$$

n = no. of observations

p = no. of predictors / regressors.

$n = 100$

$SS_{tot} = 100$

① 1 - predictor.

$SS_{res} = 40$

$R^2 = 1 - \frac{40}{100} = 0.6$

$Adj.R^2 = 1 - (1 - 0.6) \times \frac{99}{98}$

$= 1 - 0.4 \times \frac{99}{98}$

$= 1 - 0.4041$

$\boxed{Adj.R^2 = 0.5959.}$ $+ve$

② 2 - predictors

$SS_{res} = 38$

$\Rightarrow R^2 = 1 - \frac{38}{100} = 0.62$

$Adj.R^2 = 1 - (1 - 0.62) \times \frac{99}{97}$

$\boxed{Adj.R^2 = 0.6122}$

③ 3 - predictors $n-p-1$ $100-3-1$

$SS_{res} = 37.5$

$\Rightarrow R^2 = 1 - \frac{37.5}{100} = 1 - 0.375 = 0.625$

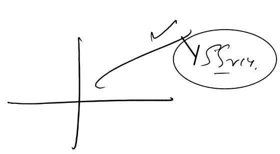
$Adj\ R^2 = 1 - (1 - 0.625) \times \frac{99}{96}$

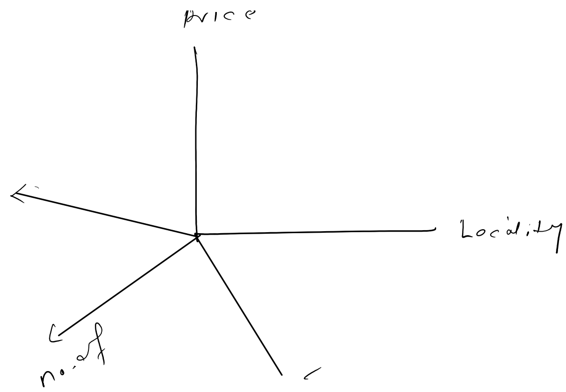
$= 1 - (0.375) \times \frac{99}{96}$

$\boxed{Adj.R^2 = 0.6131}$

X

Price




price
Locality
no. of