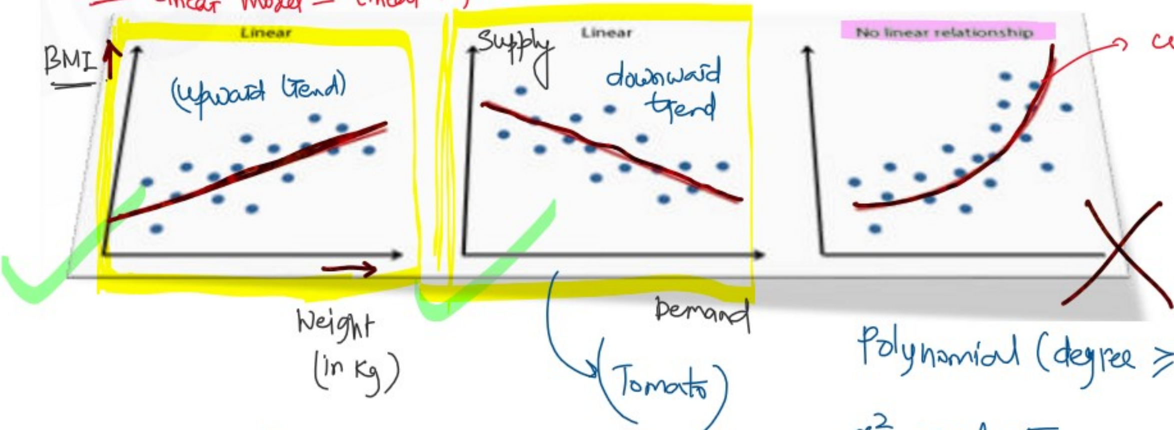
Linear Regression

11 February 2024 22:15

- data points
- Linear model — linear regression



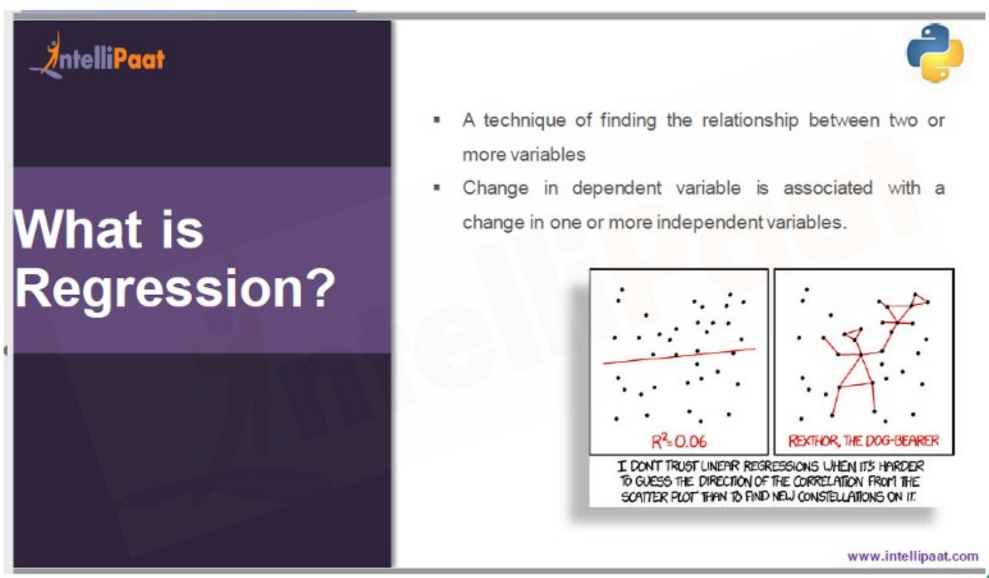
Polynomial (degree ≥ 2)

x^2: quadratic

x^3: cubic

x^4: Bi-quadratic

BMI = f(weight, height)



$y = f(x_1, x_2, x_3 \dots x_n)$

- a functional mapping
- to find the relationship between y and two or more x variables.

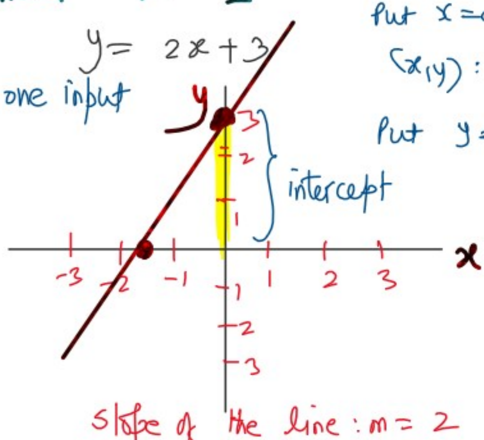
$\boxed{y = mx + c}$ (slope point form)

slope (m), independent (x), (target) (y), constant (intercept) (c)

[Simple Linear Regression]

it has only one input

$y = 2x + 3$



Put $x = 0$, $y = 2 \times 0 + 3 = 3$

$(x, y) : (0, 3)$

Put $y = 0$, $x = -\frac{3}{2} = -1.5$

$(x, y) : (-\frac{3}{2}, 0)$

slope of the line : $m = 2$

intercept of the line : $c = 3$

$y = mx + c$ → slope

(2)

$y = 2x + 3$

slope of the line : $m = 2$
intercept of the line : $c = 3$

$$\boxed{\frac{dy}{dx} = 2}$$

First order derivative is slope which is 2.

(Multiple Linear Regression)

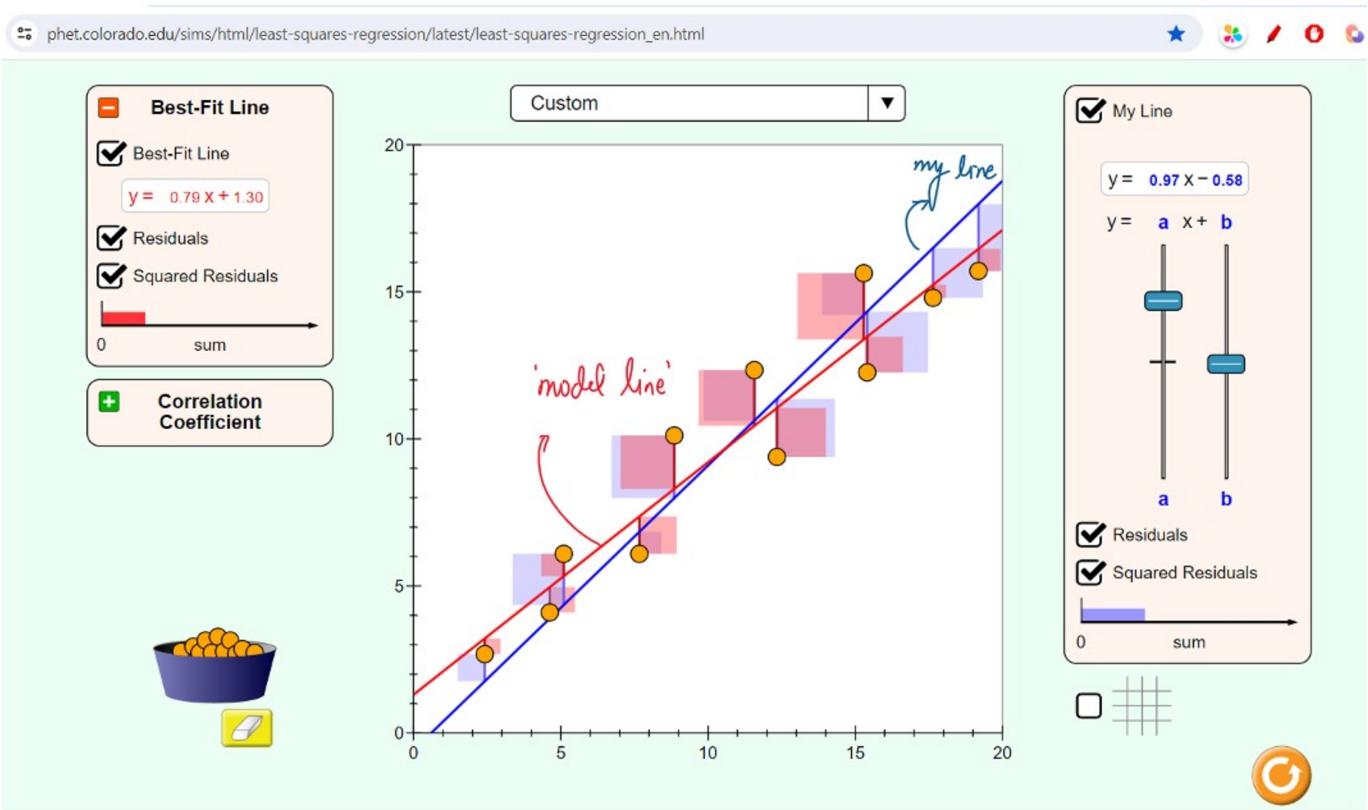
$$y = 3 + 2x_1 + 5x_2 + 3.7x_3 - 0.8x_4$$

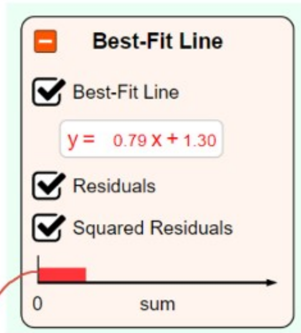
Notations

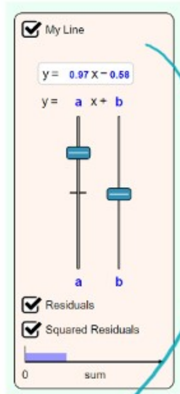
x : input variable | features | independent | predictor

y : output variable | target | dependent | response.

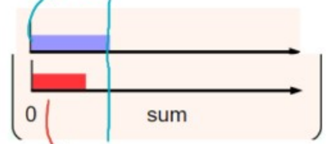
Intuition behind Linear Regression:







My line

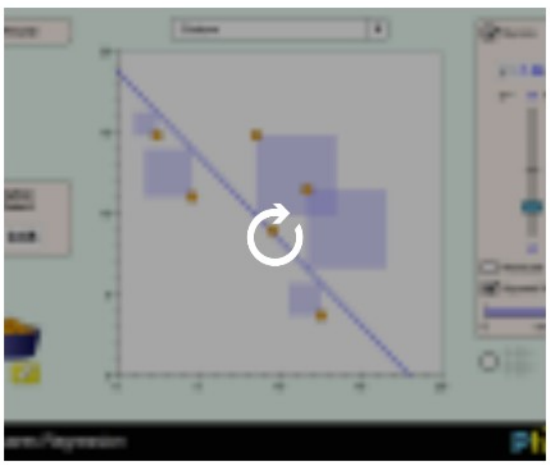


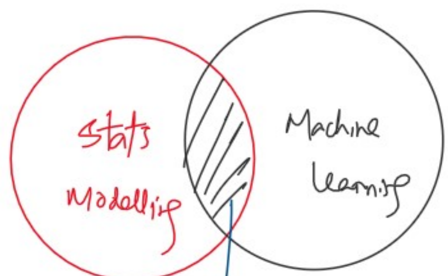
Linear Regression Model

$y = 0.79x + \boxed{1.30}$

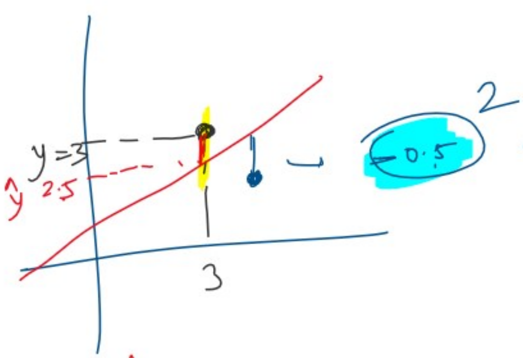
$y = 0.97x \boxed{-0.58}$

[Least-Squares Regression]





Linear Regression



^{2}

$\hat{y} = 2.5$

$y_{actual} = 3$

$delta = (3 - 2.5) = 0.5$

$square\ delta = (0.5)^2 = 0.25$

$+$ 0.25

square delta $= (0.5)^2 = 0.25$ + 0.25

sum of square residues

$y = mx + c$ (linear)

$m = -2$

$2x + y = 18$

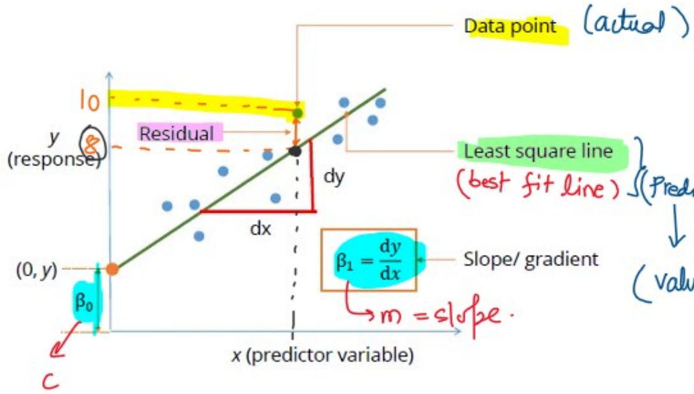
$y = -2x + 18$ (18)

Terminologies in Linear Regression

$c = \beta_0 \quad m = \text{slope} = \beta_1$

• : data points

— Linear Regression line



$y = c + mx +$

$y \quad \hat{y}$ Residual

$$y = \underbrace{\beta_0 + \beta_1 x}_{\text{Predicted value}} + u$$

Actual value

(Predicted) (value or data point)

actual value = 10 (Predicted − Actual)

Predicted value = 8 $= (8 - 10) = -2$

$$\boxed{y} = \boxed{\beta_0 + \beta_1 x} \qquad y_{actual} = \hat{y}_{predicted}$$

$y = c + mx$ $\quad y$

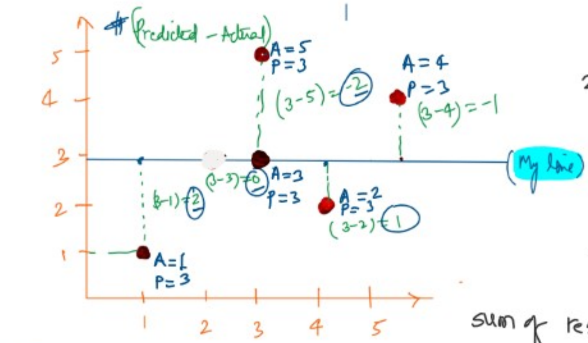
$\underbrace{\hat{\beta}_0 + \hat{\beta}_1 x} = \hat{y}$

y : actual value 10

$\hat{y}$: Predicted value 8

residue $= (\hat{y} - y) = 8 - 10 = -2$

* Issue with just doing — sum of residues



Residues

1st Point: (P−A) = 3−1 = 2

2nd Point: = 3−3 = 0

3rd Point = 3−5 = −2

4th Point = 3−2 = 1

5th Point = 3−4 = −1

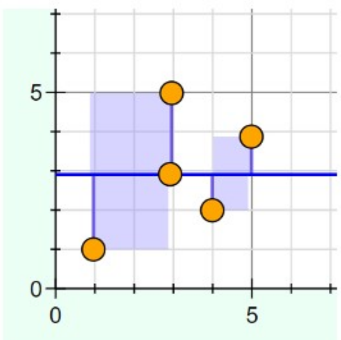
sum of residues = $\cancel{2}+0-\cancel{2}+\cancel{1}-\cancel{1}$

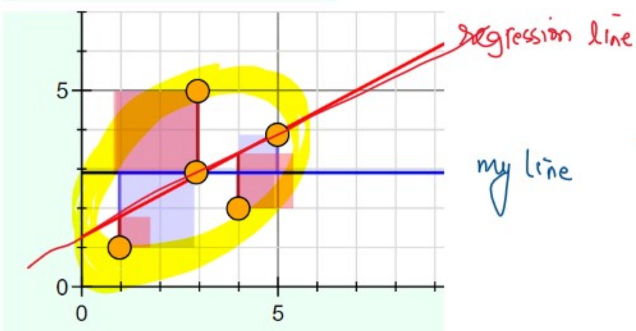
= $\underline{\underline{0}}$.

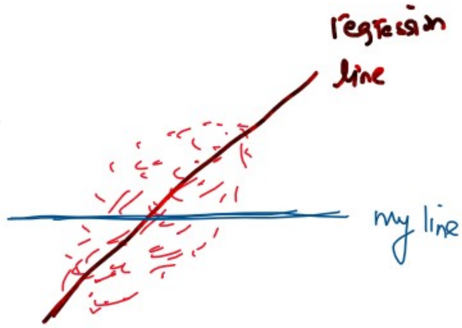
① sum of residues = $\cancel{2}+0\cancel{-2}\cancel{+1}\cancel{-1}=\underline{\underline{0}}$.

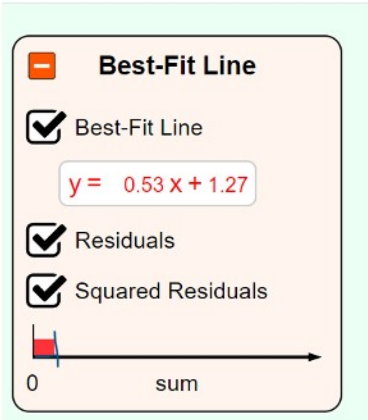
(zero sum of residues)

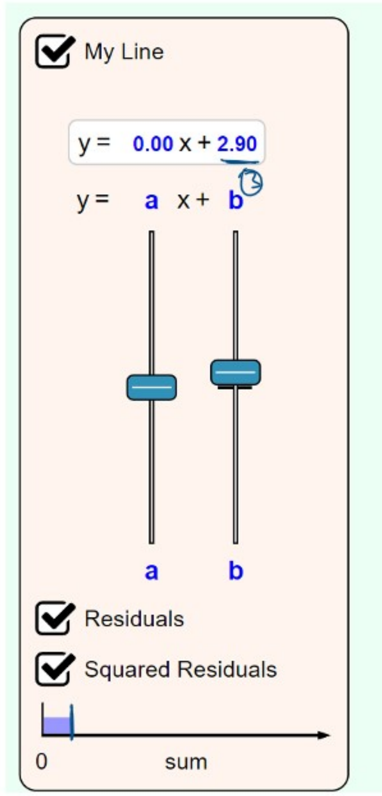
Observation: Residues nullify each other and may not be truly representing the residues.









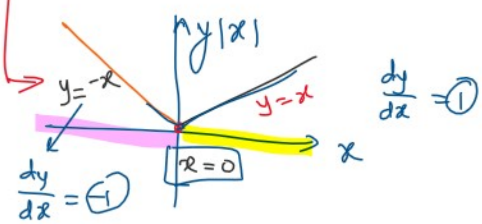


(Modulus / Absolute Function)

$$|x| = \begin{cases} x & x \geq 0 \\ -x & x < 0 \end{cases}$$

sum of residues = $2 + 0 |-2| + |1| + |1|$
$= \underline{\underline{0}}$.

$2 + 0 + 2 + 1 + 1 = 6$



|x| is not differentiable at $x = 0$

Link:

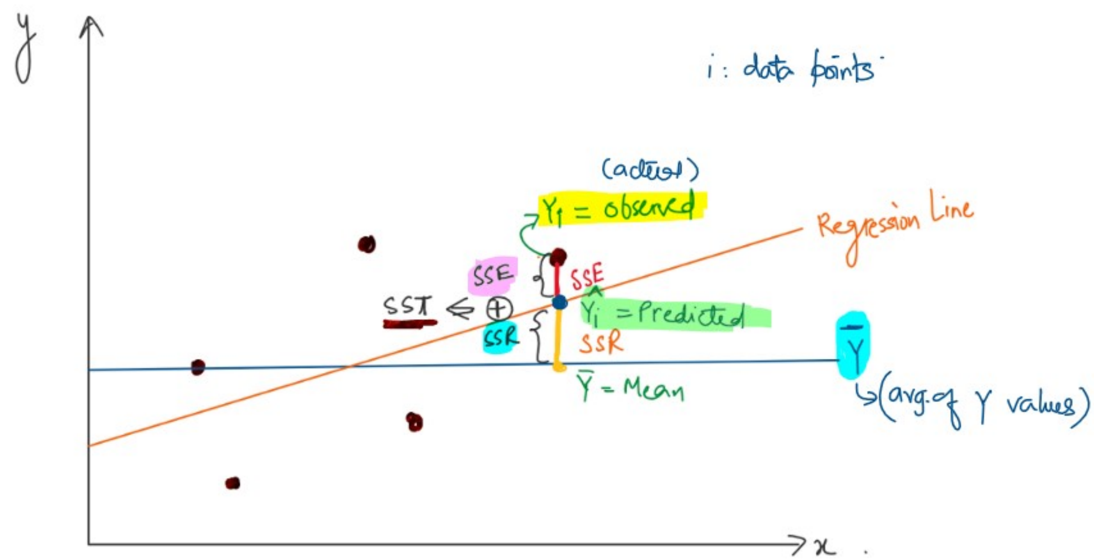
minimize the modulus of residues

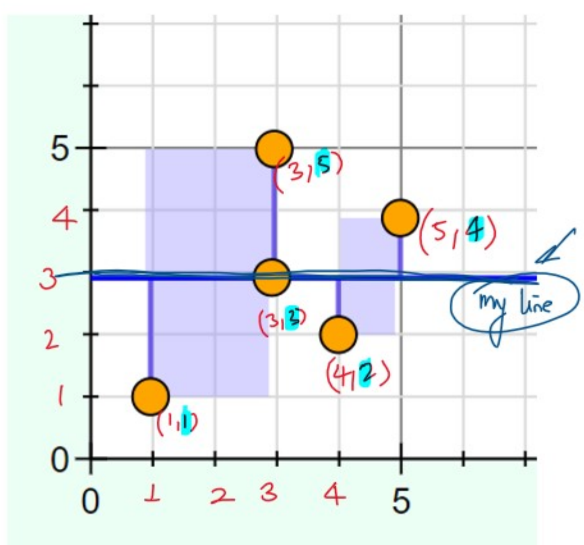
$\Downarrow$

since modulus function is non-differentiable and maxima/minima concept can't be used

Hence, let us introduce the concept of [OLS — ordinary least square]

SSE, SSR and SST





$$\bar{Y} = \frac{1+3+5+2+4}{5}$$

$$\bar{Y} = \frac{15}{5} = 3$$

$$\bar{X} = \frac{1+3+3+5+4}{5}$$

$$\bar{X} = \frac{16}{5} = 3.33$$

SSE : sum of squares error

SSR: sum of squares regression.

SST: sum of squares total

sum of squares error (residues) : SSE

- it is the sum of the squared differences between the observed value (actual) Y_i and predicted value $\hat{Y}_i$

- SSE shows the unexplained variance by regression.

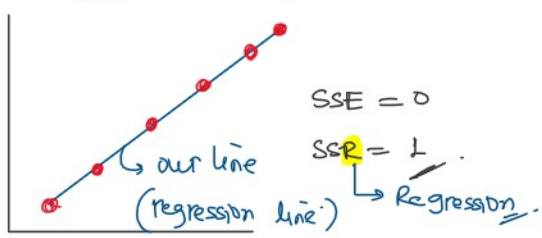
$$\boxed{SSE = \sum_{i=1}^{n} \left(\hat{Y}_i - Y_i\right)^2}$$

sum of squared error

sum of squares regression (SSR)

- it is the sum of squared differences between predicted value $(\hat{Y}_i)$ and the mean of the dependent value $(\bar{Y})$

- SSR shows the explained variance by regression
- it is a measure that describes how well our line fits the data.



$$\boxed{SSR = \sum_{i=1}^{n} \left(\hat{Y}_i - \bar{Y}\right)^2}$$

(regression)

sum of squares total (SST)

- it is the squared differences between observed dependent variable and its mean ($\bar{y}$) (y_{actual})

$$SST/TSS = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

total sum of squares

- it is a measure of total variability of the dataset

$$\boxed{SST = SSR + SSE}$$

Total variability of the dataset (SST) = Variability explained by the regression line (SSR) + unexplained variability (SSE)

100% = (80%) (accuracy) + (20%) (error)

Total variance = 1000

Total error = 100

Total regressed value = 900

Accuracy = $\frac{\cancel{900}}{\cancel{1000}} \times 100$ = 90%