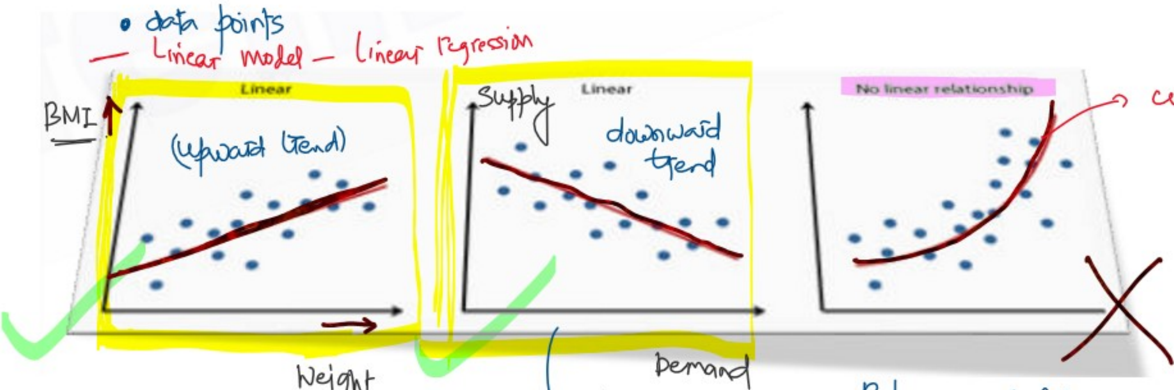
Linear Regression

11 February 2024 22:15

• data points

— Linear model — linear regression



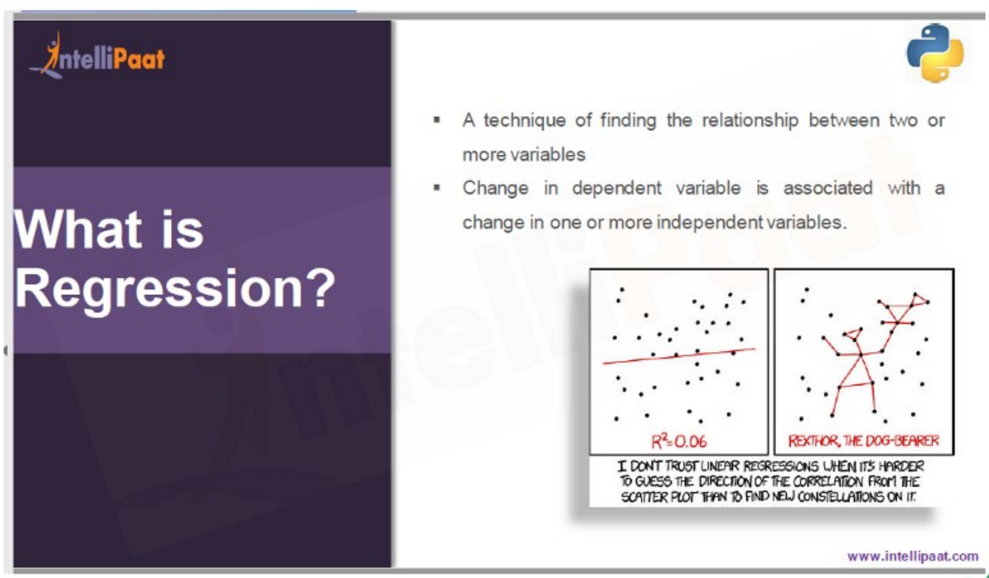
Polynomial (degree ≥2)

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

independent

(slope point form)

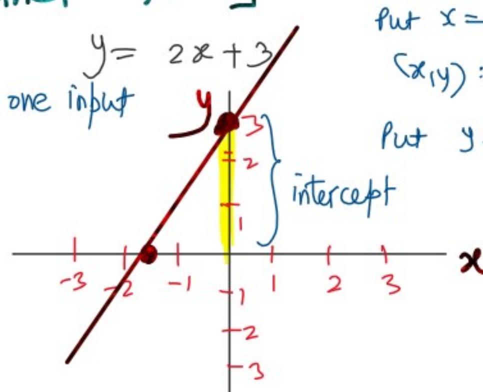
(target)

constant (intercept)

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

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

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)

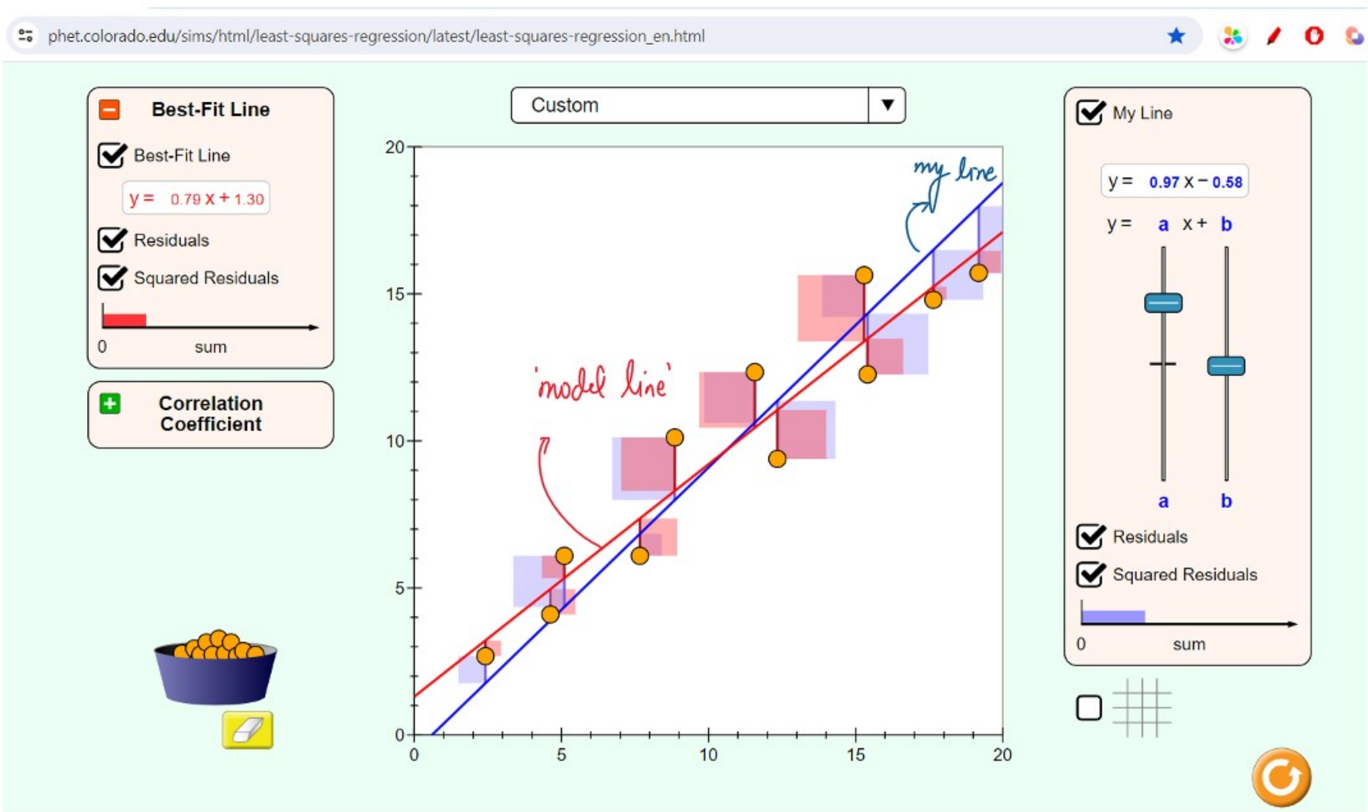
$$y = 3 + 2x_1 + 5x_2 + 3.7\,x_3 - 0.8\,x_4$$

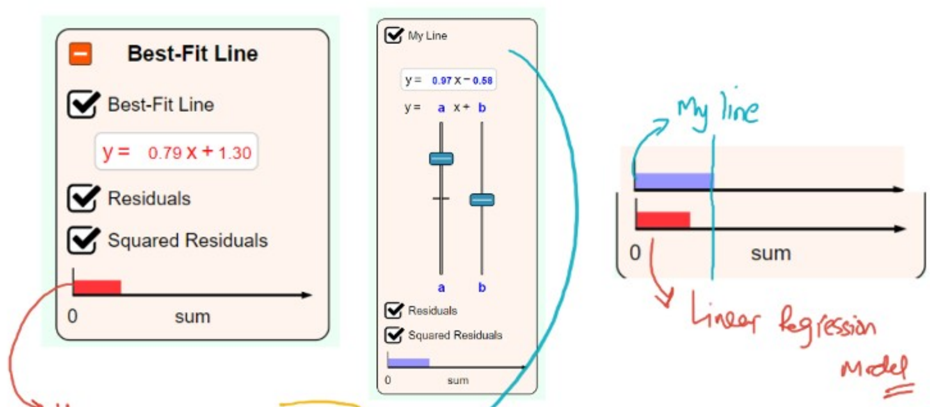
Notations

x : input variable | features | independent | predictor

y : output variable | target | dependent | response.

Intuition behind Linear Regression:

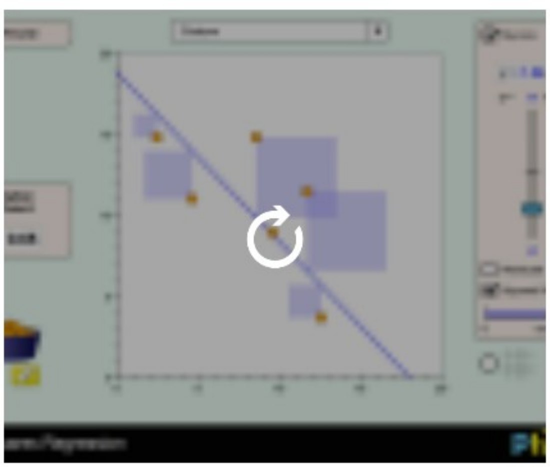


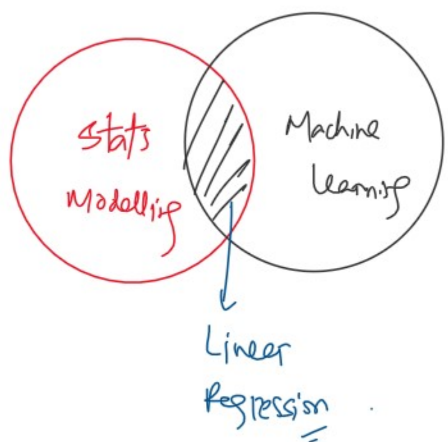


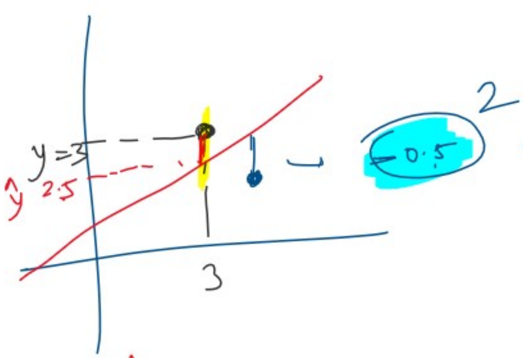
$y = 0.79x + \boxed{1.30}$

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

[Least-Squares Regression](#)







$\hat{y} = 2.5$

$y_{actual} = 3$

$\text{delta} = (3 - 2.5) = 0.5$

$\text{square delta} = (0.5)^2 = 0.25$ $+$ 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

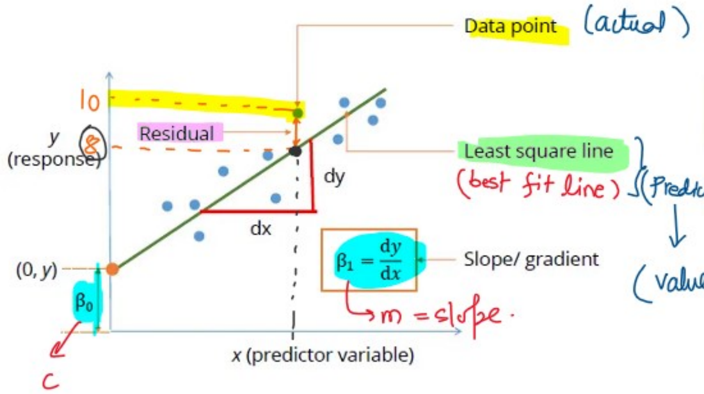
Terminologies in Linear Regression

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

• : data points

— Linear Regression line

$y = c + mx +$



y $\hat{y}$ Residual

$$y = \beta_0 + \beta_1 x + \underline{u}$$

Actual value

Predicted value

actual value = 10 (Predicted − Actual)

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

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

$y = c + mx$

y

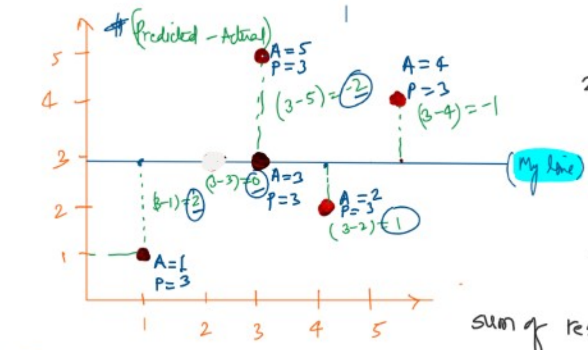
$\underbrace{\hat{\beta}_0 + \hat{\beta}_1 x}$ $= \hat{y}$

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

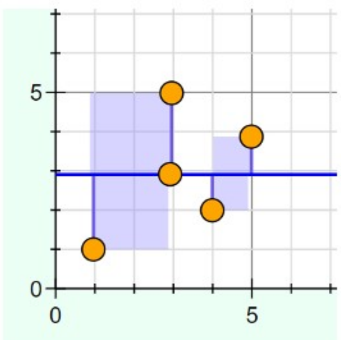
5th Point = 3−4 = −1

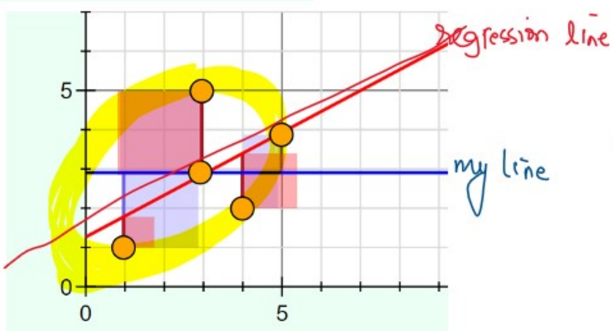
sum of residues = $\cancel{2} + 0 - \cancel{2} + \cancel{1} - \cancel{1}$
= 0

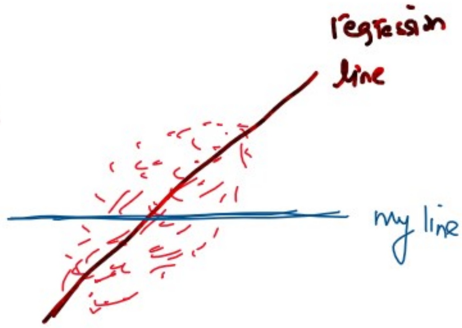
(1) sum of residues = $\cancel{2} + 0 \cancel{-2} \cancel{+1} \cancel{-1} = 0$

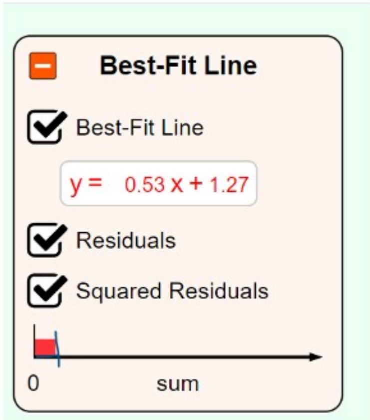
(zero sum of residues)

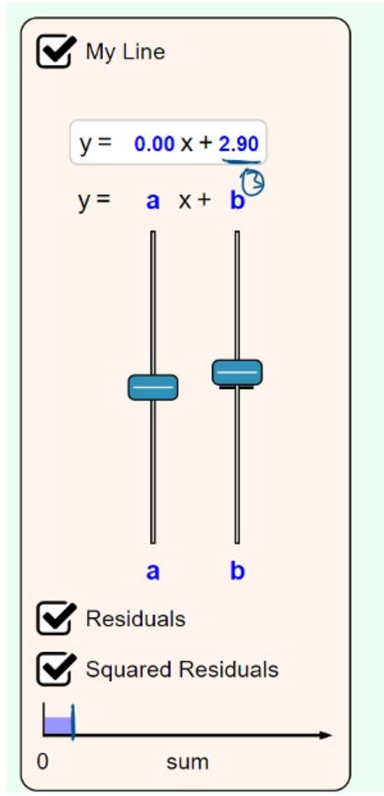
Observation: Residues nullify each other and may not be truly representing the residues.











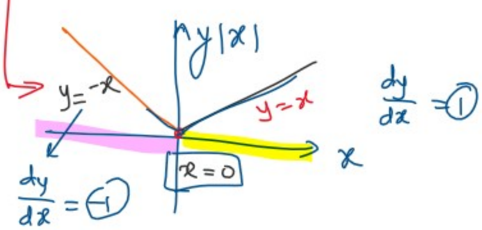
(Modulus / Absolute Function)

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

sum of residues $= 2 + 0 |-2| |-1| + |1|$

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

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



minimize the modulus of residues

⇓

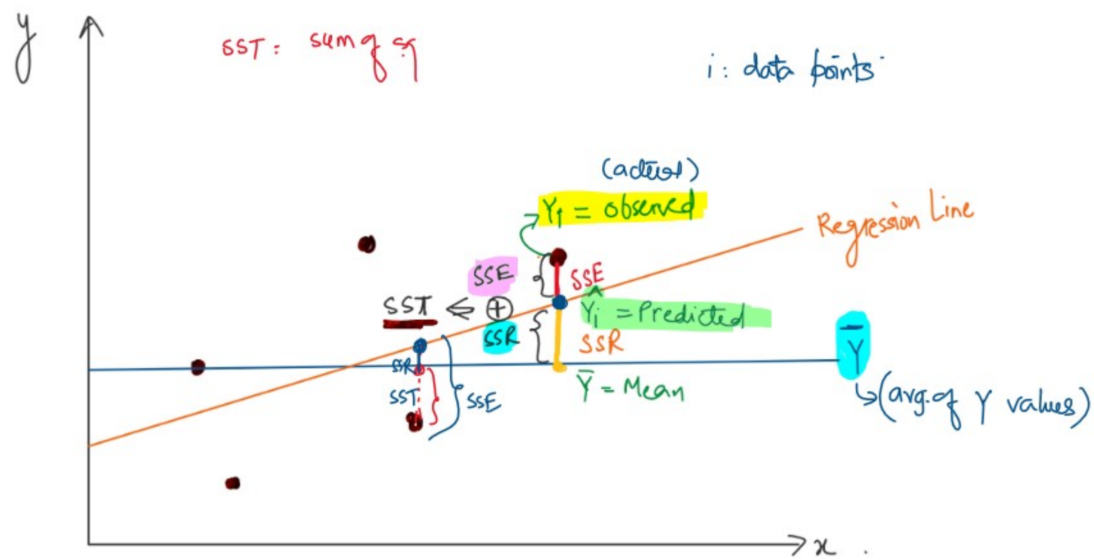
$|x|$ is not differentiable at $x = 0$

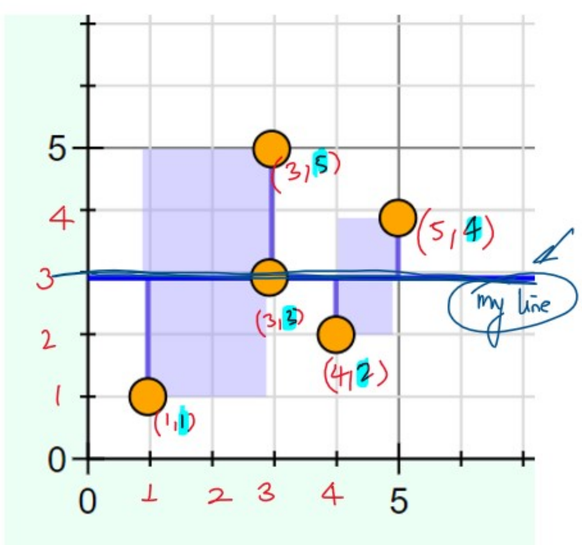
Link:

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.

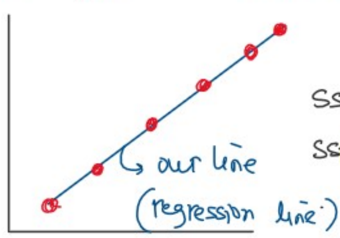
$$\boxed{SSE = \sum_{i=1}^{n} \left(\hat{Y}_i - Y_i\right)^2}$$

sum of $\boxed{\text{squared}}$ 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.



SSE = 0

SSR = 1

$\hookrightarrow$ Regression.

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

Sum of squares total (SST)

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

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

n ← all the data points

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 = $\frac{9\cancel{0}0}{1\cancel{0}00} \times 100$ = 90%

Derivation of linear regression equation

SLR: simple linear regression.

$y = mx + c$

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

$\hat{y} = a + bx + u$

$$y = a + \underbrace{bx}_{} + u$$

$$\hat{Y} \longleftrightarrow Y_i$$

delta

$$\boxed{u = \sum_{i=1}^{n} \underbrace{(Y_i - \hat{Y}_i)}^{2}} \rightarrow \text{(Minimize)} \downarrow\downarrow\downarrow\downarrow\downarrow$$

sum of squared residues | sum of squared error

SSE

$$u = \sum_{i=1}^{n \rightarrow \text{no. of rows}} (Y_i - a - bx_i)^2$$

i represents the each data points

Using differential calculus – Maxima & minima, let us partially differentiate w.r.t. 'a' and 'b' respectively.

Y_i actual

$\hat{Y}_i = a + bx_i$

$$u = \sum_{i=1}^{n} (Y_i - a - bx_i)^2$$

$y = x^2$

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

$y = x^n$

$\frac{dy}{dx} = n \cdot x^{n-1}$

objective: To calculate equation for the intercept 'a' and slope 'b'

'a' $\downarrow$ $\hat{\beta}_0$, 'b' $\downarrow$ $\hat{\beta}_1$

Class 12th (NCERT)

$z = f(x, y)$

$z = \boxed{x^2} y$

a) $\frac{\partial z}{\partial x} = y\left(\frac{\partial}{\partial x}(x^2)\right)$

$= (2xy)$

b) $\frac{\partial z}{\partial y} = ?\ x^2$

$\frac{\partial z}{\partial y} = x^2 \cdot \left(\frac{\partial y}{\partial y}\right) = x^2$

Let us differentiate 'u' partially w.r.t 'a'

$$\boxed{\frac{\partial u}{\partial a} = 0} \rightarrow \text{(critical points)}$$

$\hookrightarrow$ refer $\rightarrow$ Maxima | Minima

$$u = \sum_{i=1}^{n} \boxed{(Y_i - a - bX_i)}^2$$

concept of chain rule:

working of chain rule

$y = \boxed{x}^{2}$

$\frac{dy}{dx} = 2\,x^{2-1} \times \frac{dx}{dx}$

constant

constant

∂u $\quad n \quad (2-1)$

concept of chain rule:

$$\frac{dy}{dx} = 2x^{2-1} \times \frac{dx}{dx}$$

$$= 2x^1 \cdot 1$$

$$= 2x$$

$$\frac{\partial y}{\partial a} = \sum_{i=1}^{n} 2\underbrace{(Y_i - a - bX_i)}^{2-1} \cdot \frac{\partial}{\partial a}(\cancel{Y_i}^{0} - a - \cancel{bX_i}^{0})$$

constant

constant

$$\frac{\partial y}{\partial a} = \sum_{i=1}^{n} 2(Y_i - a - bX_i) \times \left(-\frac{\partial a}{\partial a}\right)$$

$$(Y_i - a - bX_i) \Leftrightarrow x$$

$$\frac{\partial y}{\partial a} = -2 \sum_{i=1}^{n} (Y_i - a - bX_i) = 0$$

$$\Rightarrow \sum_{i=1}^{n} Y_i - \sum_{i=1}^{n} a - \sum_{i=1}^{n} bX_i = 0 \quad \text{—— (1)}$$

$$\bar{X} = \frac{\sum_{i=1}^{n} x_i}{n} \qquad \left[\Rightarrow \sum_{i=1}^{n} x_i = n \cdot \bar{X}\right]$$

Divide equation (1) by 'n'

$$\Rightarrow \frac{\sum_{i=1}^{n} Y_i}{n} - \frac{\sum_{i=1}^{n} a}{n} - \frac{\sum_{i=1}^{n} bX_i}{n} = 0$$

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

$$\sum_{i=1}^{n} a = an$$

$$\Rightarrow \bar{Y} - \frac{\cancel{n}a}{\cancel{n}} - b\frac{\sum_{i=1}^{n} x_i}{n} = 0$$

$$\Rightarrow \bar{Y} - a - b\bar{X} = 0$$

$$\bar{Y} - a - b\bar{X} = 0$$

$$\boxed{\Rightarrow a = \bar{Y} - b\bar{X}}$$

Second condition

$$\frac{\partial y}{\partial b} = 0$$

$$\frac{\partial u}{\partial b} = 0$$

$$u = \sum_{i=1}^{n} \boxed{(Y_i - a - bX_i)}^2$$

$$\frac{\partial u}{\partial b} = \sum_{i=1}^{n} 2(Y_i - a - bx_i) \times \frac{\partial}{\partial b}(\cancelto{0}{Y_i} - \cancelto{0}{a} - bx_i)$$

$$\frac{\partial u}{\partial b} = 2\sum_{i=1}^{n}(Y_i - a - bx_i) * (-x_i) = 0$$

$$\frac{\partial u}{\partial b} = -2\sum_{i=1}^{n}(Y_i - a - bx_i)\underline{x_i} = 0$$

$$\Rightarrow \sum_{i=1}^{n}(x_iY_i - ax_i - bx_i^2) = 0$$

Let us substitute 'a' with $\bar{Y} - b\bar{X}$

$$\Rightarrow \sum_{i=1}^{n}\left[x_iY_i - (\bar{Y} - b\bar{X})x_i - bx_i^2\right] = 0$$

$$\Rightarrow \sum_{i=1}^{n}(Y_ix_i - \bar{Y}x_i + b\bar{X}x_i - bx_i^2) = 0$$

$$\Rightarrow \sum_{i=1}^{n}(Y_ix_i - \bar{Y}x_i) + \sum_{i=1}^{n} b(\bar{X}x_i - x_i^2) = 0$$

$$\Rightarrow \sum_{i=1}^{n}(Y_ix_i - \bar{Y}x_i) = -\sum_{i=1}^{n} b(\bar{X}x_i - x_i^2)$$

$$\Rightarrow \sum_{i=1}^{n}(Y_ix_i - \bar{Y}x_i) = b\sum_{i=1}^{n}(x_i^2 - \bar{X}x_i)$$

$$\sum_{i=1}^{n}(Y_ix_i - \bar{Y}x_i)$$

$$\Rightarrow b = \frac{\sum_{i=1} (Y_i x_i - \bar{Y} x_i)}{\sum_{i=1}^{n} (x_i^2 - \bar{x} x_i)}$$

(sum product of X & Y)

$$\Rightarrow b = \frac{\sum_{i=1}^{n} x_i Y_i - \bar{Y} \sum_{i=1}^{n} x_j}{\sum_{i=1}^{n} x_i^2 - \bar{x} \sum_{i=1}^{n} x_i}$$

$$\sum_{i=1}^{n} x_i = \boxed{n\bar{x}}$$

$$\Rightarrow b = \frac{\sum_{i=1}^{n} x_i Y_i - n\bar{x}\bar{Y}}{\sum_{i=1}^{n} x_i^2 - n(\bar{x})^2}$$

Intuitively using two expressions:

$$\sum_{i=1}^{n} (\bar{x}^2 - x_i \bar{x}) = 0 \qquad \text{lly} \qquad \sum_{i=1}^{n} (\bar{Y}^2 - Y_i \bar{Y}) = 0$$

$$= \sum_{i=1}^{n} \bar{x}^2 - \sum_{i=1}^{n} x_i \bar{x}$$

$$= \bar{x}^2 \sum_{i=1}^{n} 1 - \bar{x} \sum_{i=1}^{n} x_i$$

$$= \cancel{n\bar{x}^2} - \cancel{\bar{x} \cdot n\bar{x}} = 0.$$

$$\bar{x}^2 \times n$$

*

$$\boxed{\Rightarrow b = \frac{\sum_{i=1}^{n} x_i Y_i - n\bar{x}\bar{Y}}{\sum^{n} x_i^2 - n(\bar{x})^2}}$$

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

$$\Rightarrow b = \frac{\sum_{i=1}^{n} (Y_i x_i - \bar{Y} x_i) + 0}{\sum_{i=1}^{n} (x_i^2 - \bar{x} x_i) + 0}$$

$$\sum \bar{x}\bar{Y} - \sum Y_i \bar{x}$$

$$\bar{x}\bar{Y}\sum_{i=1}^{n} 1 - \bar{x}\sum_{i=1}^{n} Y_i$$

$$\cancel{n\bar{x}\bar{Y}} - \cancel{\bar{x} \cdot n\bar{Y}} = 0$$

$$\Rightarrow b = \frac{\sum_{i=1}^{n} (Y_i x_i - \bar{Y} x_i) + \sum_{i=1}^{n} (\bar{x}\bar{Y} - Y_i \bar{x})}{\sum_{i=1}^{n} (x_i^2 - \bar{x} x_i) + \sum_{i=1}^{n} (\bar{x}^2 - x_i \bar{x})}$$

$$\Rightarrow b = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(Y_i - \bar{Y}) / n}{\sum_{i=1}^{n} (x_i - \bar{x})^2 / n}$$

$\longrightarrow$ covariance

$\longrightarrow$ variance

$$\boxed{b = \frac{\mathrm{cov}(x, Y)}{\mathrm{var}(x)}}$$

$$\mathrm{var}(x) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}$$

$$\mathrm{cov}(x, Y) = \frac{\sum_{i=1}^{n} (x_i - \bar{x}) * (Y_i - \bar{Y})}{n}$$

$$Y = f(x)$$

$Y = f(x)$

~~Var(x) ??~~

$\boxed{\operatorname{cov}(X,Y)}$

$$b = \frac{\operatorname{cov}(X,Y)}{\operatorname{var}(x)}.$$

***

$$\boxed{\begin{array}{c} a = \bar{Y} - b\bar{X} \\ \hline b = \dfrac{\operatorname{cov}(X,Y)}{\operatorname{var}(x)} \end{array}}$$

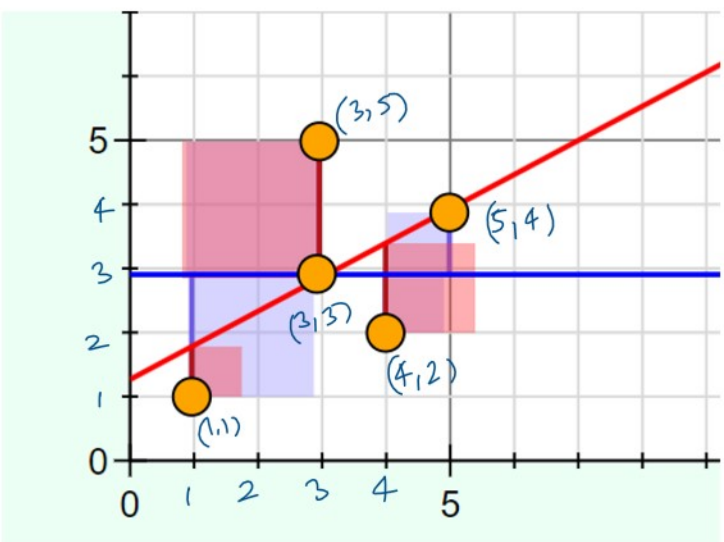
$Y = f(x)$

$\uparrow \qquad \uparrow$

supply = f(demand) ↑↑↑

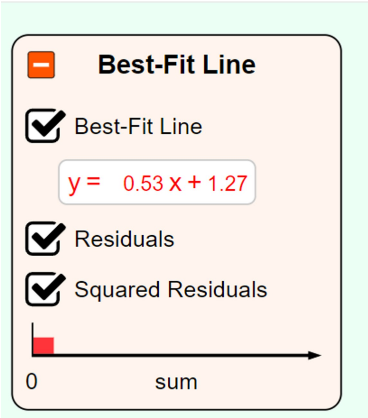
↓ out of stock

Note: 'a' is constant (Y-intercept) is such that the line (LR model) must go through the $\underbrace{\text{mean of 'x'}}_{\bar{X}}$ and $\underbrace{\text{mean of 'Y'}}_{\bar{Y}}$



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

$\bar{Y} = 3.$

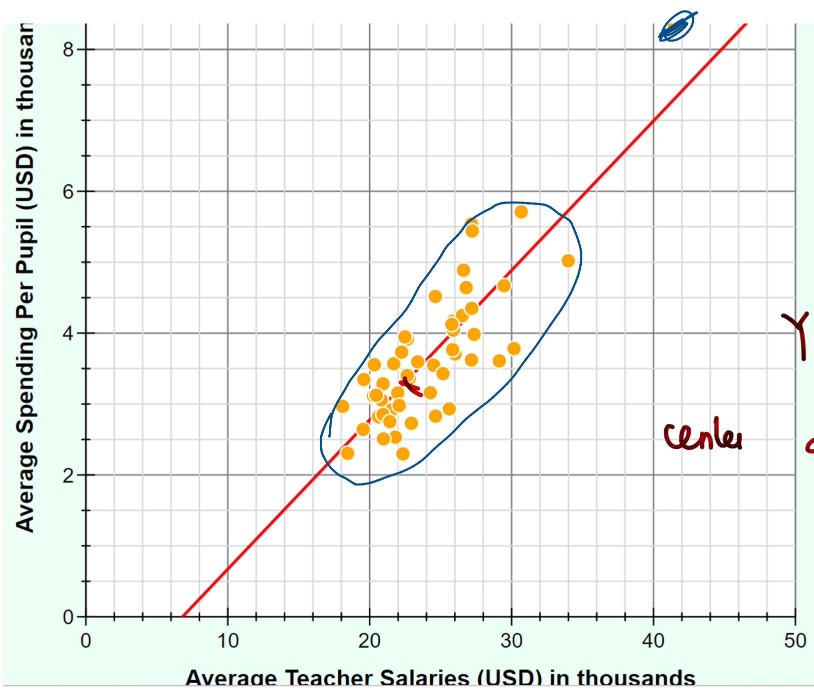


$y = 0.53x + 1.27$

$\left(\bar{x} = \frac{10}{3} , \bar{y} = 3\right)$

$3 = 0.53 \times \frac{10}{3} + 1.27$

.53*10/3=1.7667
1.7667+1.27=3.0367



Y

center

d l

ad