Linear Regression Recall...

① Machine learning

↳ the machine learn a task.

The overall Essence → Learning

Operating a Lift.

↳ Rules

↳ No prior data

Self-driving Car.

↳ Press accelor

↳ apply brake

↳ Prior data

↳ More & differn Scenarios.

Empirical learning/ Machine Learning

If you dont have data → No performance.

① Statistics ⇒ understand the data

↳ phenomenon that have generated the data

↳ often have understanding to perform testing → different hypothesis

② psychology = understand behaviours exhibited by people.

③ Data Mining : seeks to find the

pattern inside the data that is understandable by people.

→ The Learning
- ↳ Task. (T)
- ↳ Data (D) → ML algorithm (Experience E)

New Data (N) ← Perform task prediction. (Evaluate)

Algorithm + Data ⇒ Model ↓ Task.

→ Set of Instructions
↳ can also be mathematical

Child reaches → 5 years of age } 1,00,000 hrs of data

Speech Recognition
↓
Automatic Speech Recognition } Metric

MS → 16000

compare / evaluate my model

metric

← WER | 3/10

1. Task
2. Data
3. fit the Data with an algorithm
4. Evaluate the model (predictions)

Adding more data

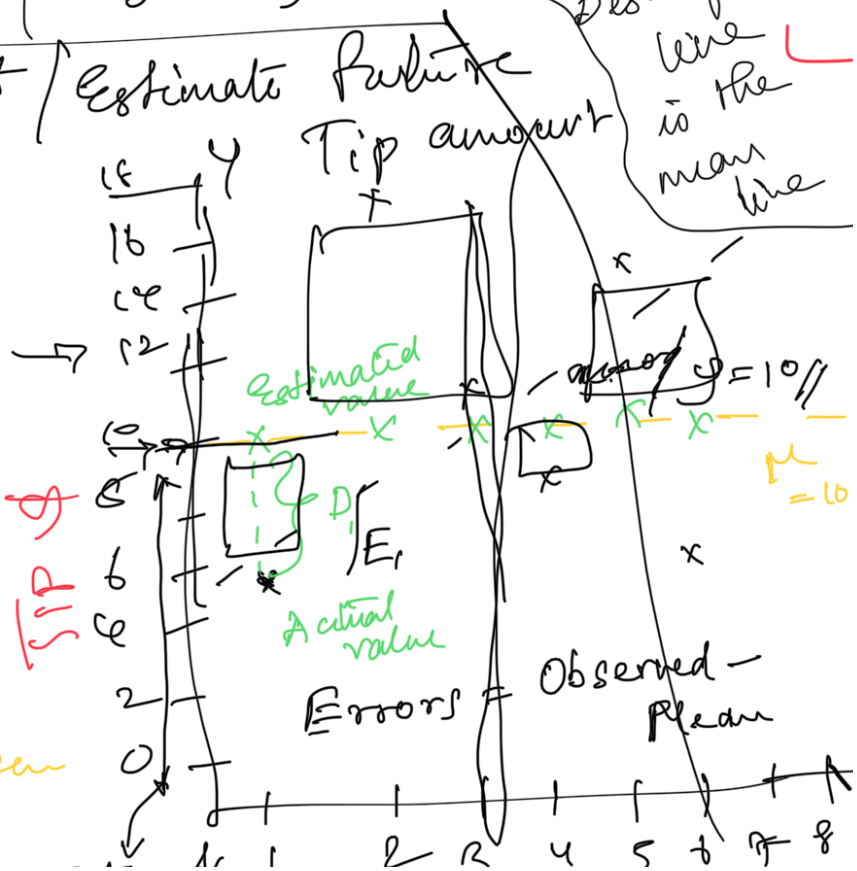
5. perform prediction in real time
↓
taking it live

Meal	Tip amount $
1	5.00
2	17.00
3	11.00
4	8.00
5	14.00
6	5.00

y

One variable

1. Task → predict / Estimate future Tip amount
2. Data
3. C.T
4. Evaluate the model



Can I Make use of Measure of C.T ??

Intercept

Meal #

Variance

Sum of Squares ( ) ⇒ Sum (observed − Mean)^2

Deviations

n

Mean Sum of Squares

n

↳ Sum of Squared Deviations | Mean Sum of Squared Deviations

Error

Errors

Before we proceed can we do better??

Benchmark. | Baseline

↳ SSB

↳ SSE

↳ SST

↳ 120

Algorithm → reduces my SSE → Computing again.

$y = 10$

rate of change = Change in Y / Change in X = 0/0

vertical line

$x = 3$

rate of change = 0/0 → ∞

(Function ⇒ ΔY → delta) = $\frac{y_2 - y_1}{x_2 - x_1}$

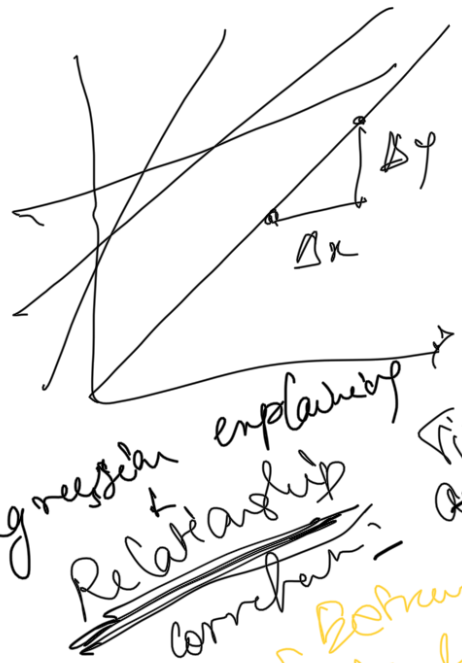
Slope = $\frac{\text{change of res}}{\Delta x}$

$y = f(\text{Slope}, \text{Intercept})$

$y = mx + c$

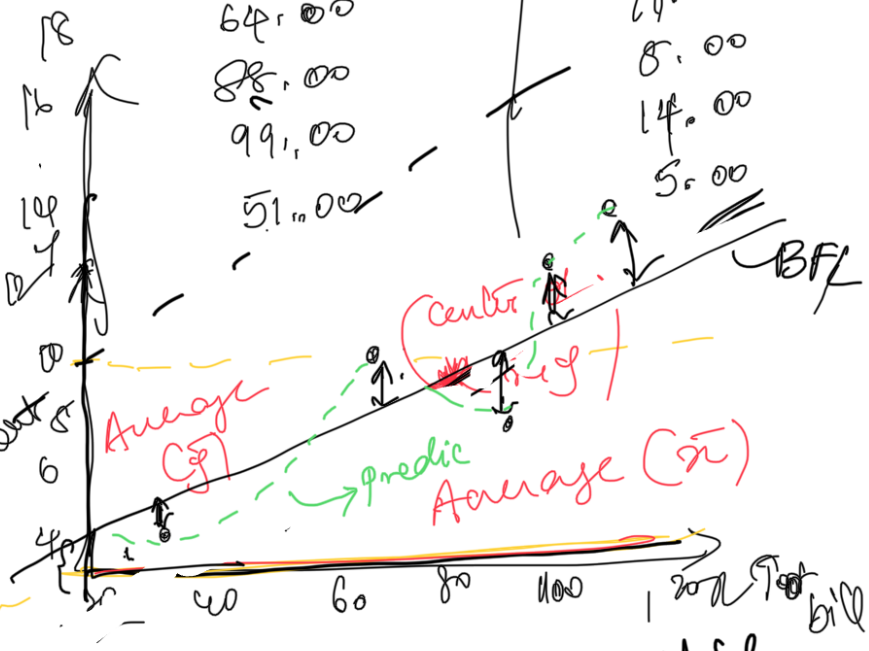
m → Slope, c → Intercept

define your target as a function of x



Regression explains the relationship between correlation

Total bill (₹)	Tip amount (₹)
34.00	5.00
108.00	17.00
64.00	14.00
88.00	8.00
99.00	14.00
51.00	5.00



Brute force method

Best fit line → Line which has min Error.

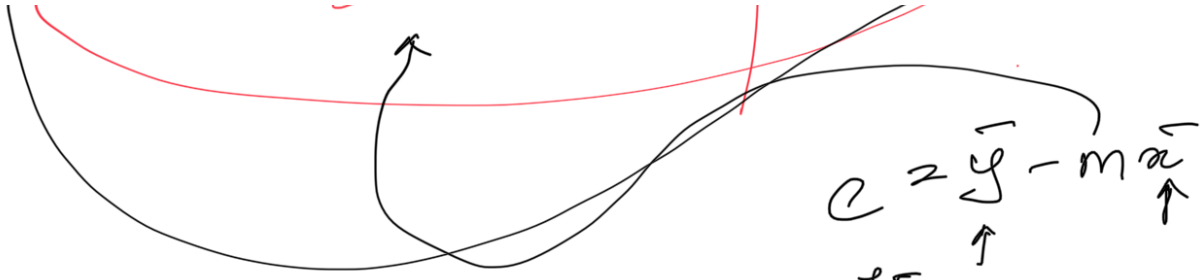
$m = \frac{\Delta y}{\Delta x}$

$y = mx + c$

$c = y - mx$

Slope $(m) = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$

$c = \bar{y} - m\bar{x}$



$c = \bar{y} - m\bar{x}$

$SSE = 30.075$

~~SSE~~ $= SST = 120$

$120 \longrightarrow 30.075$ { new Regression line

$120 - 30.075 = 89.925$ { old mean line

$SST = 120$

$SSE = 30.075$

$SSR = 89.925$

120

Total of the Error is Explained by my regression line

Goodness of fit

Coefficient of Determination $= r^2 = \frac{SSR}{SST}$ → how good is my New/Reg. line

$\frac{120}{120} = 1$

$r^2 = 1$ Best Model

89.9

Assumptions for Linear Regression.

1. Linear Relationship bet X & Y.
2. Data should have little to no Multicollinearity — Write up on why/how is bad?

Multicollinearity

Several Independent variable(X) are correlated among themselves

2 variables are perfectly co-linear +1/-1

affect the statistical inference bet. X → Y.

Compute coefficient

slope Intercept

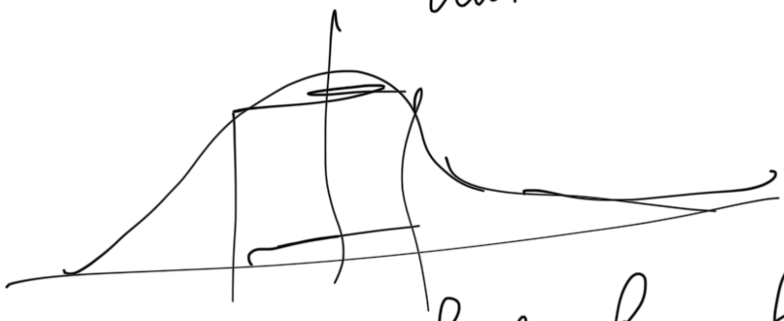
read on how/why write an article

Simple !!!

$y = m_1x_1 + m_2x_2 + m_3x_3 + \ldots + c$

Multicollinearity will inflate m, when computing this coeffic.

(3) Residuals should be normally distributed



(4) Residuals → should have homoscedasticity

Homoscedasticity → (2)

Same variance

the errors should same / Similar variances.

(2f) Write-up on Assumps of Linear Regress.