1.1 Types :

Regression
- Bayesian
- ElasticNet
- Lasso
- Ridge
- multiple Linear
- Decision Tree
- Polynomial
- Random forest
- Simple Linear
- Support Vector

——— X ——— X ———

## Simple Linear Regression

(i) Regression

→ return to a less developed state, i.e coming down to a single variable

→ techinque to investigate rel^n b/w dependant & independant variables

Eg: Historical data

| | Area | Dist | Bedrooms | Price |
|---|---|---|---|---|
| $H_1$ | | | | |
| $H_2$ | | | | |
| ⋮ | | | | |
| $H_{300}$ | | | | |

ID variable

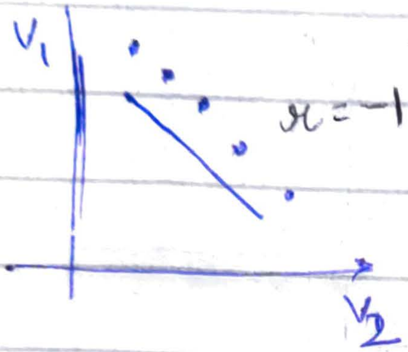
Features/ Explanatory Variables/ Independant variables

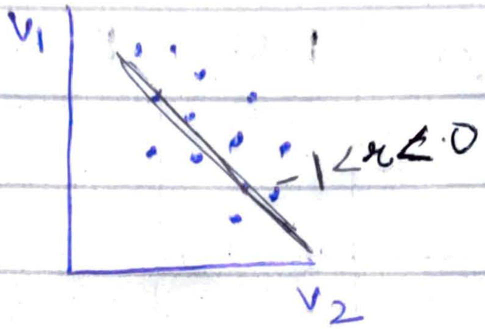
Target/Dependant vari.

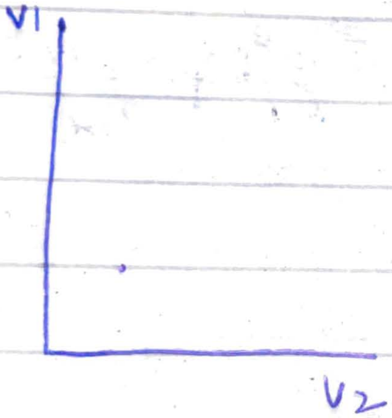
RULES:

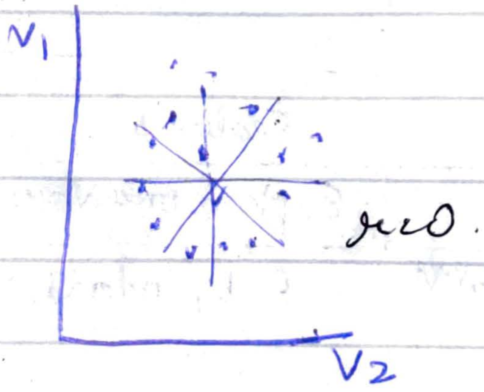
I. ID variables are not used for predictions

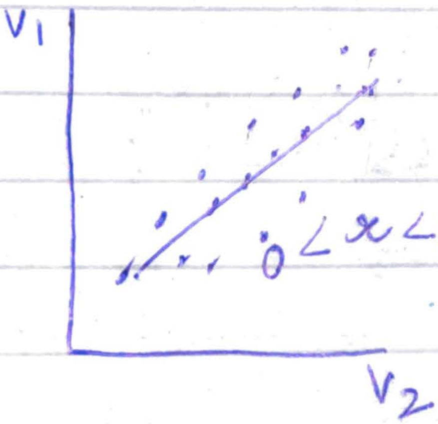
# Correlation

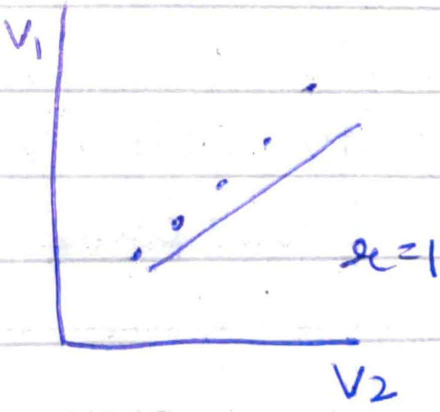


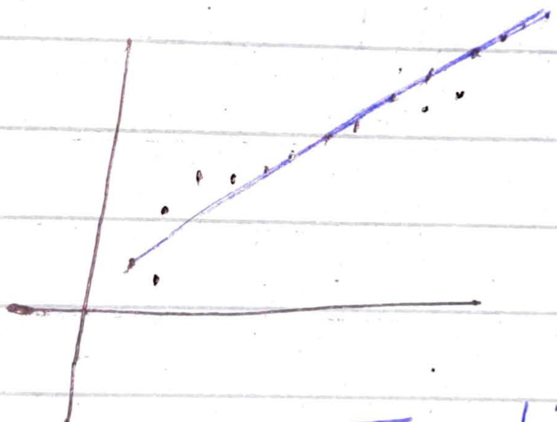



# Difference b/w

Corr = 2 variables,

reg - Variables, hypothesised,
- advance understanding of system

Ice Cream:



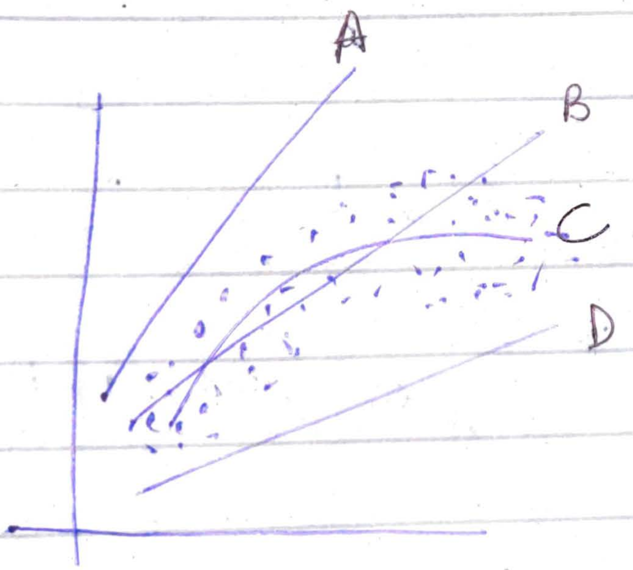
| Temp | IC | IC=2T | Error |
| --- | --- | --- | --- |
| given | given | predi. | Given -Pred |
| | | | $\boxed{\phantom{xx}}$ minimiz |

## General Hypothesis

$I \cdot C = a + b\,Temp$     a, b = unknown
(Ice Cream)

# Find a & b for which error is min.

a, b are called parameters
∴ we need to find parameters to minimize the error.



out of ABD
B is best

out of ABCD
C is best

out of B & C,
C is best as
it minimizes the error.

The best line also has errors. If we

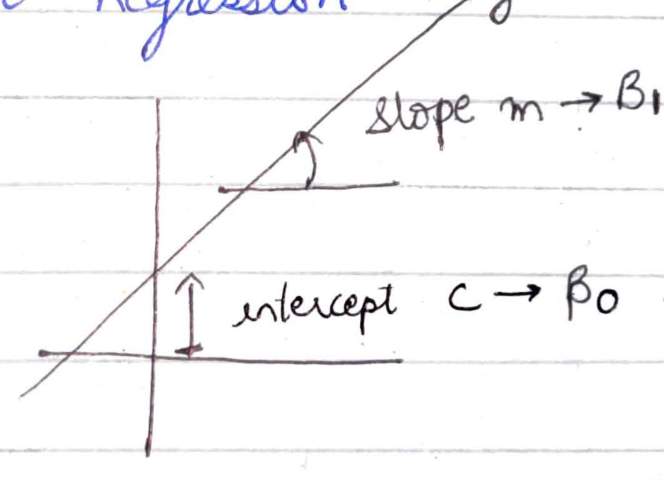
reach an error that is good enough we reach line B. Further if the error is stile significant we move further to line C.

Causation is not Regression $\rightarrow$ definate / fixed

Regression is not correlation

## Linear Regression

Line




$y = mx + c$

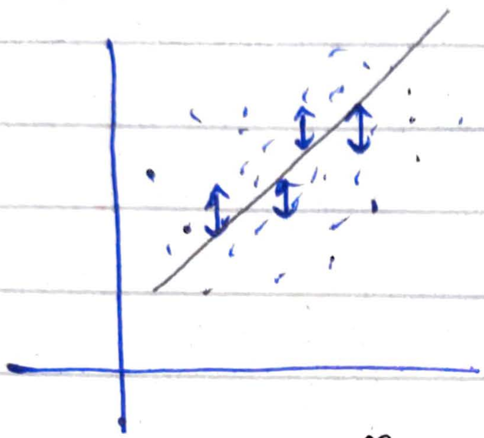
$y = B_0 + B_1 x$

n dim $\longrightarrow$ (n-1) dimensional qentity

2 d $\longrightarrow$ line (1D)

3 d $\longrightarrow$ plane (2d)

Linear regr. is a linear approach to find the relation b/w a dependant & independant variables



$\beta_0$ & $\beta_1$ such that ERROR is min.

ERROR = $Y_{actual} - Y_{predicted}$.

I $\quad \frac{1}{n} \sum_{i=1}^{n} (Y_a - Y_p)$

II $\quad \frac{1}{n} \sum_{i=1}^{n} |Y_a - Y_p|$

RMSE III $\quad \sqrt{\frac{1}{n^2} \sum_{i=1}^{n} (Y_a - Y_p)^2}$

I is not used since positive & -ve errors cancel to give a false 0 error.

II → not used as mod is used; the mod value makes the error non-differentiable; hence inc. difficulty in computⁿ.

III → used since all values are considered & hence computation also possible.
The sq. maginfies the error but the sq. root balances.

→ To make III further simple the SSR is used

SSR = Sum Square Residual.

$$SSR = \frac{1}{n} \sum_{i=1}^{n} (Y_a - Y_p)^2$$

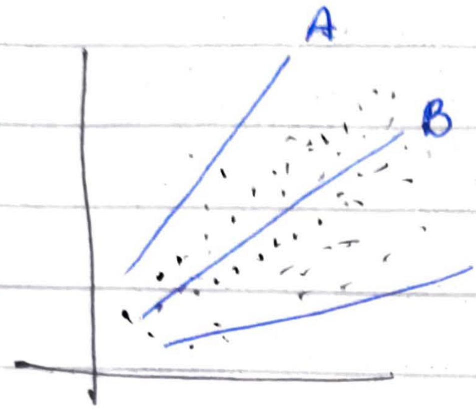
Hence $B_0$ & $B_1$ are such that SSR is min.

$$RMSE = \sqrt{SSR}$$

Model Evaluation Metrics

1) RMSE

2) R-Square




→ Least RMSE

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_p - y_a)^2}$$

$$SSR = \frac{1}{n}\sum_{i=1}^{n}(y_p - y_a)^2$$

A best fit line is not a perfect line

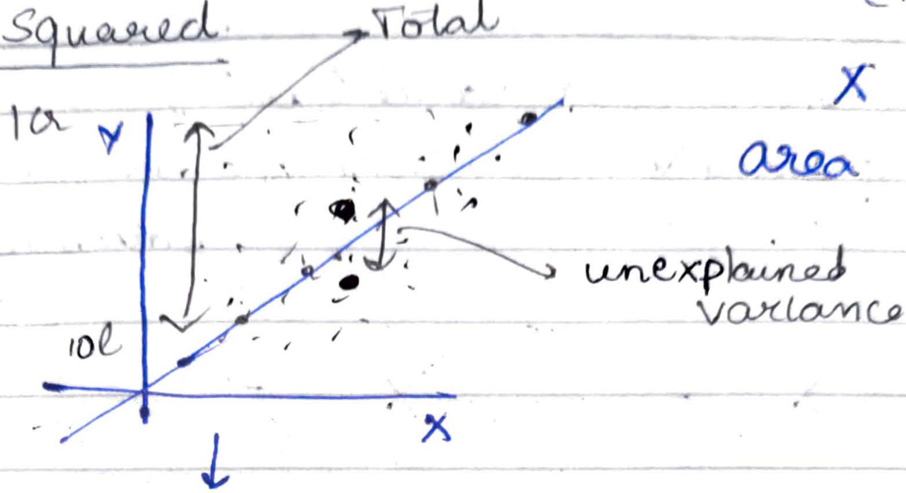
RMSE & R-sq are universal for all continous techniques.

(House)

X Y

area price

## R - Squared




area ↑ , price ↑ .

In the absence of an explanatory variable ; the spread of y is different

∴ variance of y is explaind by x

$$R^2 = \frac{\text{Explained Variance}}{\text{Total Variance}}$$

$$R^2 = \frac{\text{Total Variance} - \text{Unexp. Variance}}{\text{Total Variance}}$$

∴ $R^2$ is the % of variance explained.

Total Var = Variance (in Y)

Unexplained Var = Variance in error column

———— X ———— X ————

Multiple Dimensions

p explanatory variables

$X_1$ $X_2$ $Y$ --> p explanatory values

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots \beta_p X_p$$

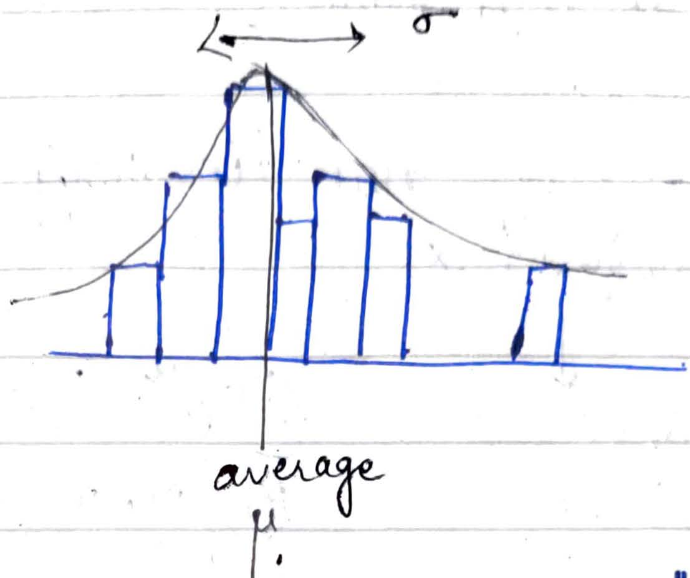
↓

"Hyper plane"

Error = || to y axis & not $\perp$r to the plane.

# ASSUMPTIONS

① Normal Distribution




$\mu \pm \sigma = 66\%$

$\mu \pm 3\sigma = 99\%$

① The target variable should be normal distributed

no skewness.

if not normal, transform y

sqrt (y)
log (y)

$x \sim \log y$

$y = e^{x}$

2) Linear relationship

→ TV & Independant Variable should be linearly related

- In case of non-linearity, we can transform x

① $x^2$
② $\sqrt{x}$
③ $\log x$
④ $e^x$
⑤ $\sin x$

4

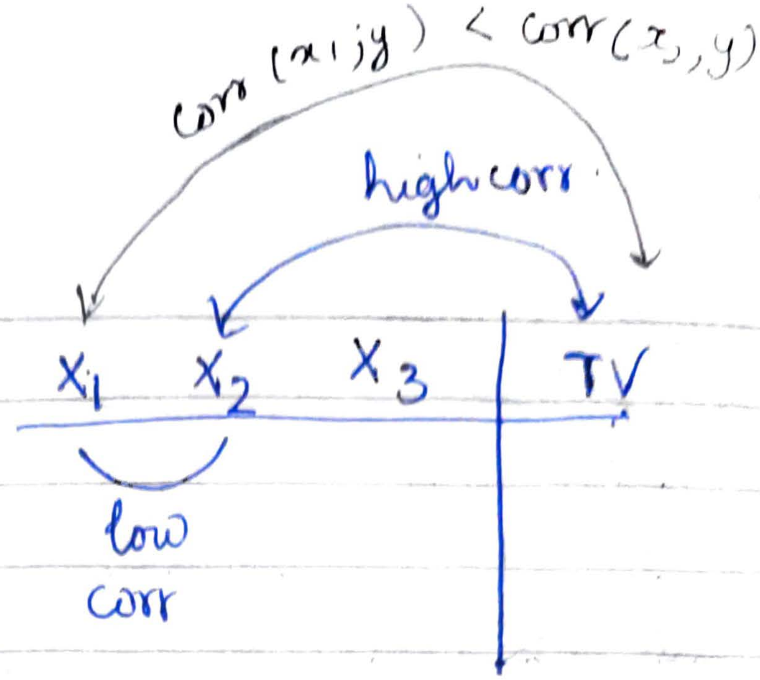
3) Independant Variables should not be correlated

"no corr" & truely independant

$x_1$ & $x_2$ → low correlation

$|r| < 0.75$

if 2 columns are highly correlated hence 1 is dropped.




$X_1$ is dropped.

even if

(4) Variance in error should be Constant → called as homoscedasticity

(1) done post analysis; as eval$^{n}$ is done on error (after pred$^{n}$)

(2) If variance in error is not constant; the model is defective. The variance in error is not constant only if the previous assumptions are violated ~~or~~ or transform of variables is not correct.