* SLR.
  - Gradient Descent
  - Evaluation metrics

Scikit Linear Regression

{ R square :=> % Y by X
{ adjusted R-square :=>

MSE
RMSE
MAE.

→ SLR → two datasets.

Agenda
- Assumption of Linear Regression
- Multiple linear Regression
- Polynomial regression.
- Lasso, ridge, Elastic Net.
- Cross validation & Hyperparameter tuning
- Logistic Regression.

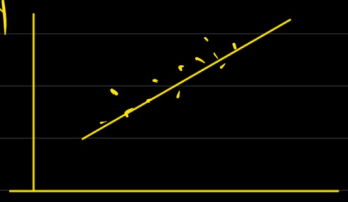
* Multiple linear Regression

Pari	# of rooms	Locality	Area of house	Price of house	% of syllabus completed	No. of hours Studied	Marks obtained
						5	58

* All the information Y is not captured by X, in order to capture more & more information we require more X.

MLR ⇒ more the 1 IV is used.

1 – IV ⇒ $y_{pred} = \theta_0 + \theta_1 x$. ⟶



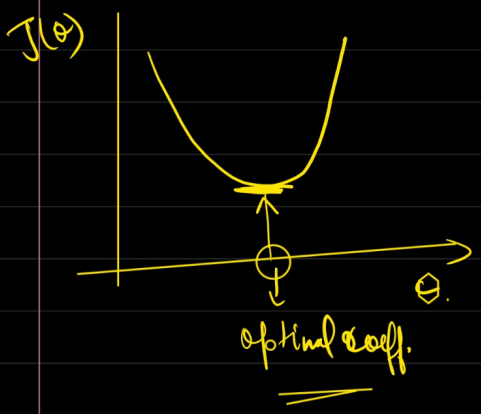
$$h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 \cdots \theta_n x_n$$

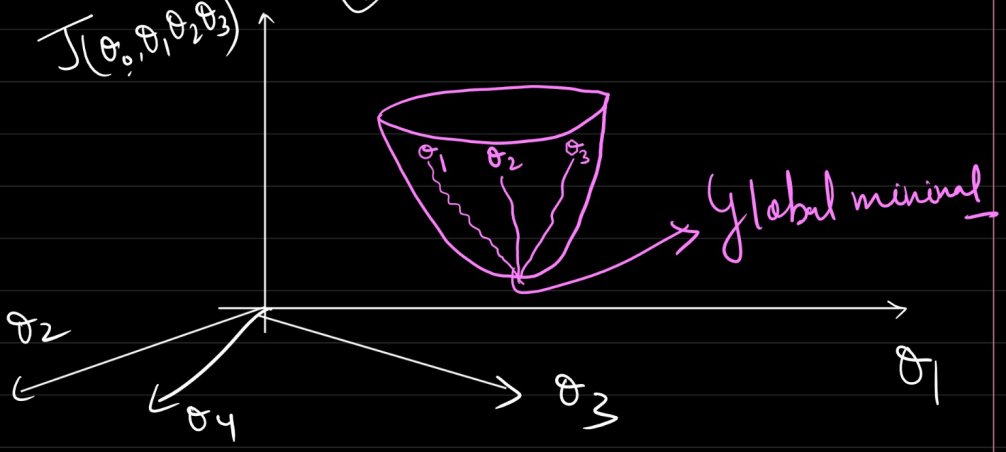
$\theta_1, \theta_2 \cdots \theta_n$ are optimal coeff.

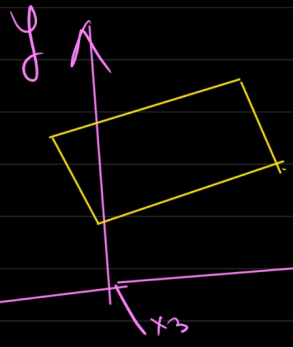
$$CF = \frac{1}{n} \sum_{i=1}^{n} (y_{act} - y_{pred})^2$$

$$(y_{act} - (\theta_0 + \theta_1 x_1 + \theta_2 x_2 \cdots \theta_n x_n))^2$$

Earlier in SLR.



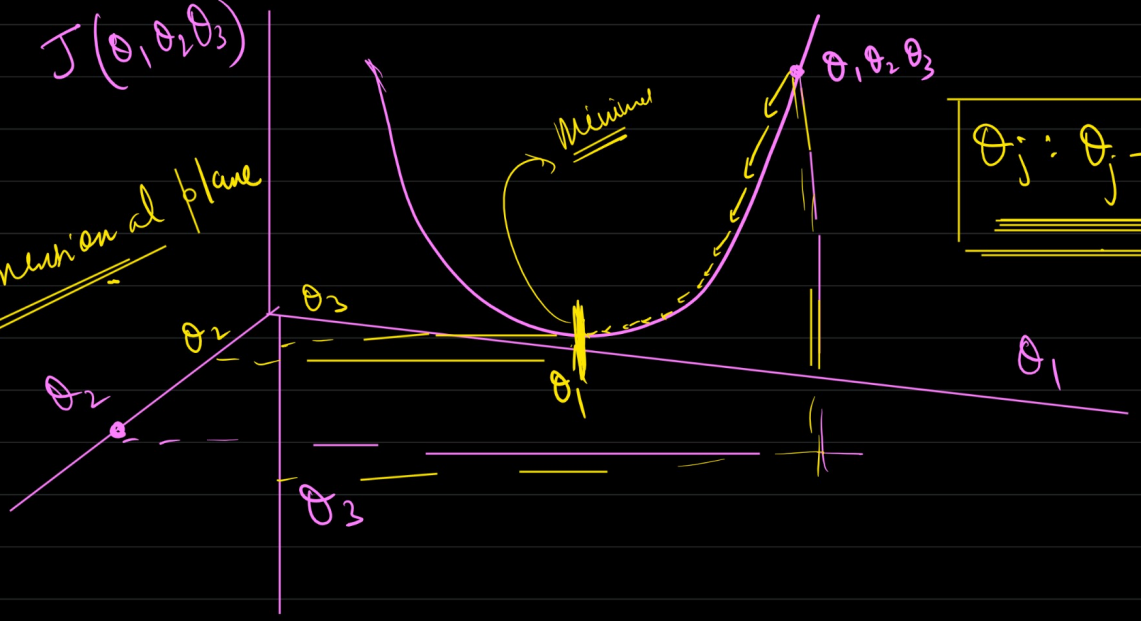




⇒ In MLR, the intercept is the point where hyperplane intersects the vertical axis (dependent variable Y)

$$h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots \theta_n x_n$$

Evaluation metrics



$$\theta_j := \theta_j - \eta \frac{\partial CF}{\partial \theta_j}$$

$$X_{fit} - x_1, x_2, x_3, x_4, x_5 \downarrow$$

$$\theta_1 \quad \theta_2 \quad \theta_3 \quad \theta_4 \quad \theta_5 \quad \theta_0$$

Q1

# of hours studied	Marks obtained
-	-
-	-
-	-
-	-
-	-

θ_0 → Avg marks obtained by a student even if he didn't studied

θ_1 → 0

$y = 50 + 2.5x$

$\Rightarrow \theta_1 \Rightarrow$ With 1 unit increase in # of hours studied, the marks increase by 2.5 units.

9:33 ↳ 9:50

Q2

Speed of car	Age
-	-
-	-
-	-
-	-
-	-

$\longrightarrow y_{Avg} = 120 + 3.2\,(Age)$.

Age = 0 → price of car is 120 Lpa

1 unit increase in X, SP of car decreases by 3.2 units

Fig

Area

θ_1

$y = 1.2 + 3.2x$

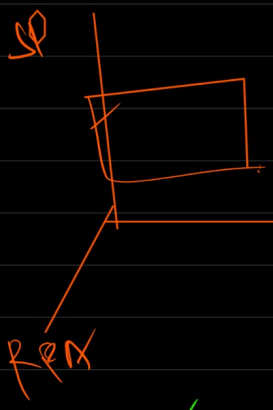
→ with 3.2 ↓ ↑ x

y ↑ 3.2 unit

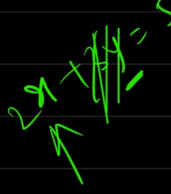
interpretation of 1.2 doesn't make

MLR

$y_{pred} = 3 - 0.52(\text{Age of car}) + 2.5\,RPM$



$\theta_1(0.52)$ → With 1 unit increase in age of car the SP of car decreases by by 0.5 units on an avg keeping RPM constant



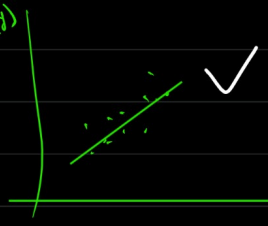
$\beta_2(2.5)$

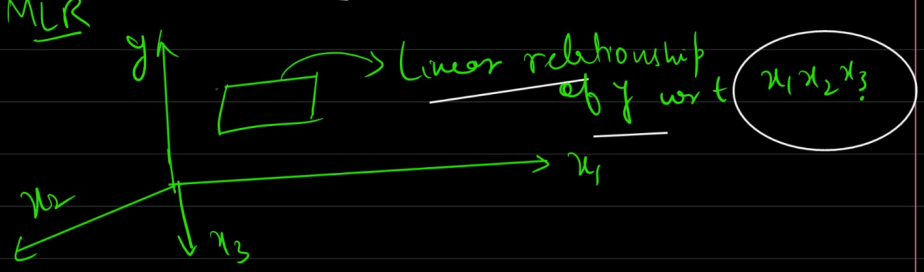
H/w
- → Sckit learn. what algorithms used
- → Explain / Unexplain Variation in Square
- → Statsmodel.
- → Feature importance in linear regression

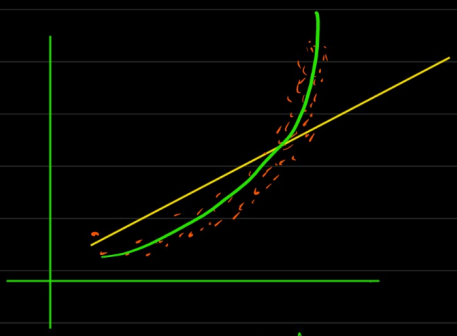
* Polynomial Regression

→ Simple linear Regression, $h_\theta(x) = \theta_0 + \theta_1 x$

for MLR : $h_\theta(x) = \theta_0 + \theta_1 x_1 + \dots + \theta_n x_n$







* Polynomial regression.

Simple LR (1 DV - 1 IV)

① Simple Polynomial regression (1 DV, 1 IV)

Polynomial degree 0, $h_\theta(x) = \theta_1 x^0$

$= \theta_1$

Polynomial degree 1: $h_\theta(x) = \theta_0 x^0 + \theta_1 x_1$ (1)

$\Rightarrow \theta_0 + \theta_1 x_1$

⇓ Simple L.R.

Polynomial degree 2

$h_\theta(x) = \theta_0 x^0 + \theta_1 x_1^1 + \theta_2 x_1^2$ (2)

$h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_1^2$

Polynomial degree 3 $h_\theta(x) = \theta_0 x^0 + \theta_1 x_1^1 + \theta_2 x_1^2 + \theta_3 x_1^3$

$\Rightarrow \theta_0 + \theta_1 x_1 + \theta_2 x_1^2 + \theta_3 x_1^3$

* As you increase the degree, you might get an overfitting model.

→ Non-linear relationship → Decision Trees

Polynomial degree 'n' = $h_\theta(x) = \theta_0 + \theta_1 x_1^1 + \theta_2 x_1^2 + \theta_3 x_1^3 + \dots \theta_n x_1^n$

for multiple X (for 3 X)

Polynomial degree 2: → x^2

$\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 +$

$\theta_4 x_1^2 + \theta_5 x_2^2 + \theta_6 x_3^2 +$

$\theta_7 x_1 x_2 + \theta_8 x_2 x_3 + \theta_9 x_3 x_1$

$x_1 x_2$ ← Cross product keeping power in mind

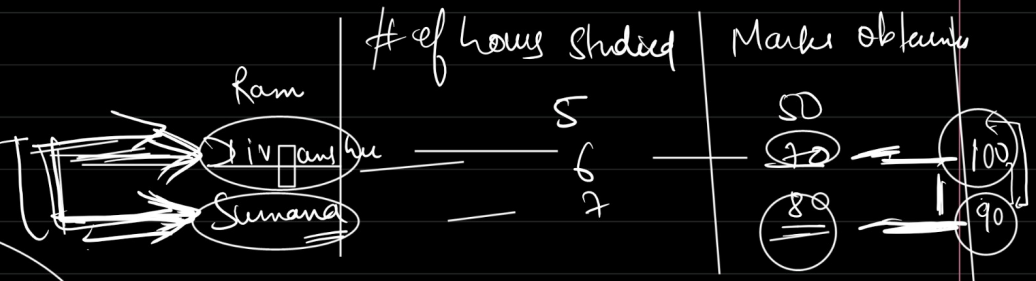
$x_1 x_2 \quad x_2 x_3 \quad x_1 x_3 \quad x_1 \quad x_2 \quad x_3$

* Assumption of Linear Regression

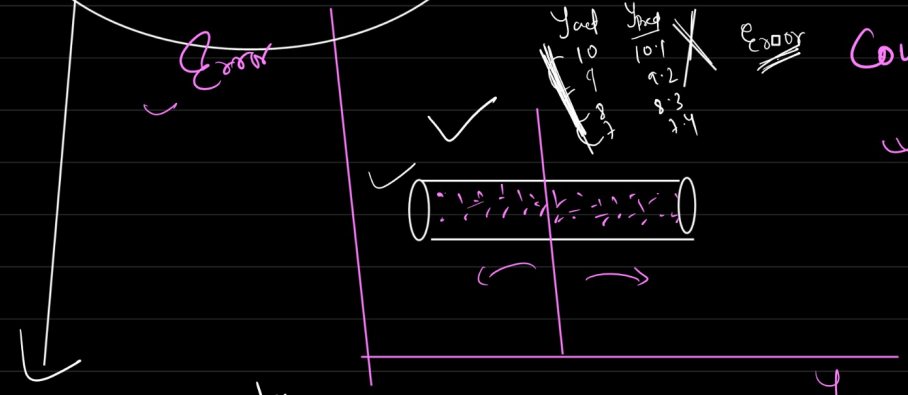
① Linearity → X and Y should have linear relationship.

② Independence → Observations (rows) are independent of each others.

→ Errors should be independent



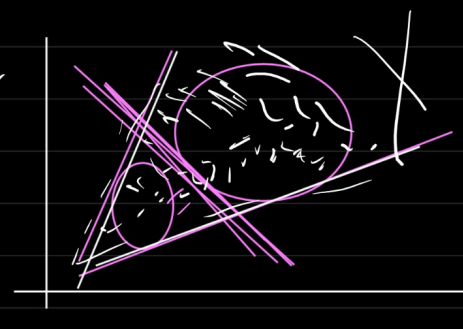
③ Homoscedasticity — Also known as Constant variance. the variance of errors are Constant.

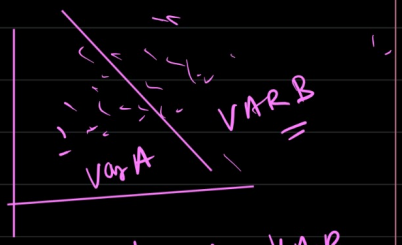


How

→ Durban-Watson test.

Heteroscedasticity





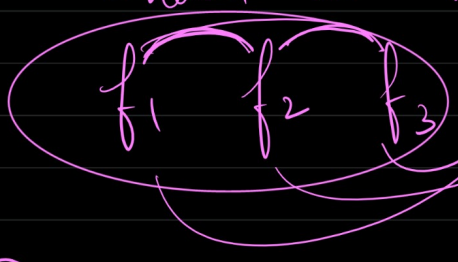
VarA = VARB

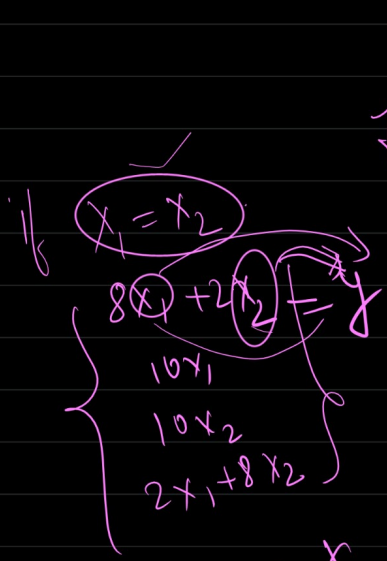
↳ Homoscedastic

④ Normality of Error

→ Errors should be normally distribute.

⑤ The feature should not be related or Should have relation





Multi − Col − linearity

Many together linear relation

$\{X_1 - X_2\}$ — Correlation

$\{X_1 - y\}$

$\{X_1 \simeq (X_2, X_3)\}$

$\{X_1 \simeq (X_2, X_3, X_4)\}$

Multi-Collinearity → where a feature exhibits a linear relationship with more than one variables.

(xy)

$\{X_1 \sim X_2\ X_3$

$X_4 \sim X_2$ — — X_5

$X_5 = +1\ X_2\ X_3$

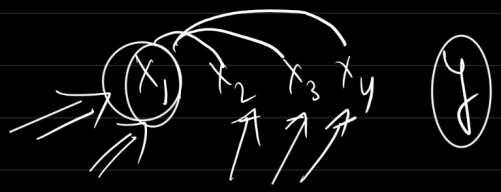
Multicollinearity

What → One feature explained by other features

Why → You can not interpret correctly what is the contribution of each individual feature w.r.t y

Severe? → No effect on predictions

Problem → interpretability
→ Computationally expensive – model training

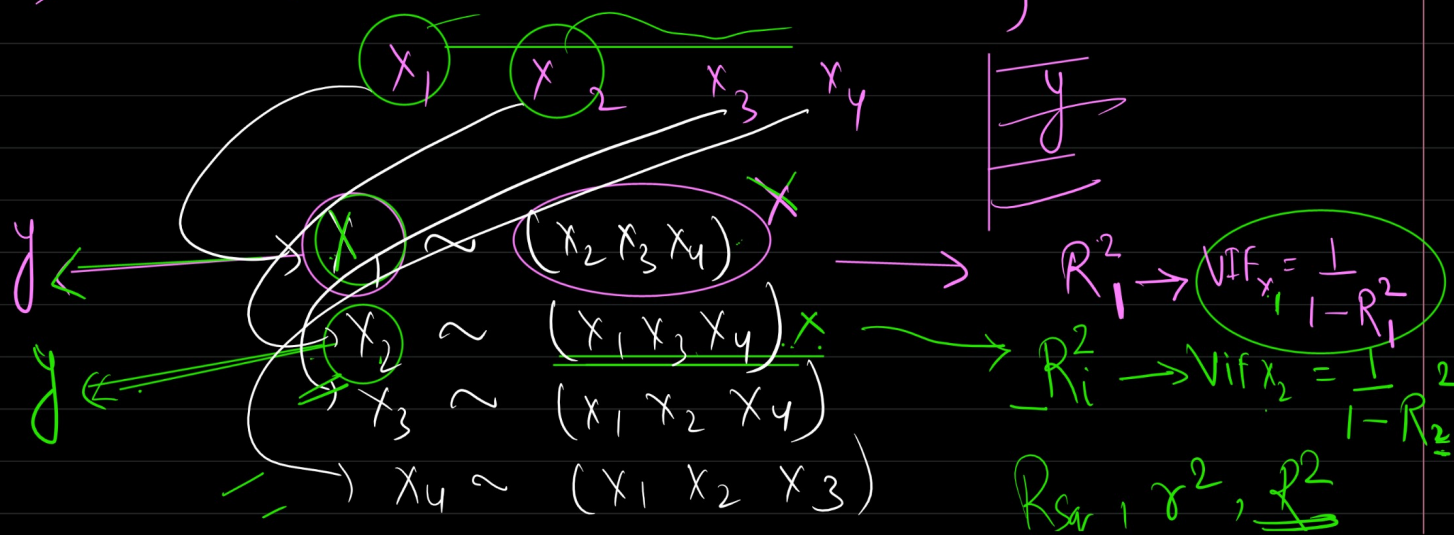


How? → VIF (Variance inflation factor)
↳ a measure of amount of multicollinearity in regression

$$VIF_i = \frac{1}{1-R_i^2}$$

→ Rsquare → % age variation in y explained by X

$$X_1 \sim (X_2\ X_3\ X_4 \ldots X_n)$$



R_{sq}, r^2, R^2

* Since VIF doesn't impact prediction, you should always ask business team before dropping any feature based on VIF

$X_1 \sim (X_2 X_3 X_4)$ — 200

$X_2 \sim (X_1 X_3 X_4)$ — 5

$X_3 \sim (X_1 X_2 X_4)$ — 8

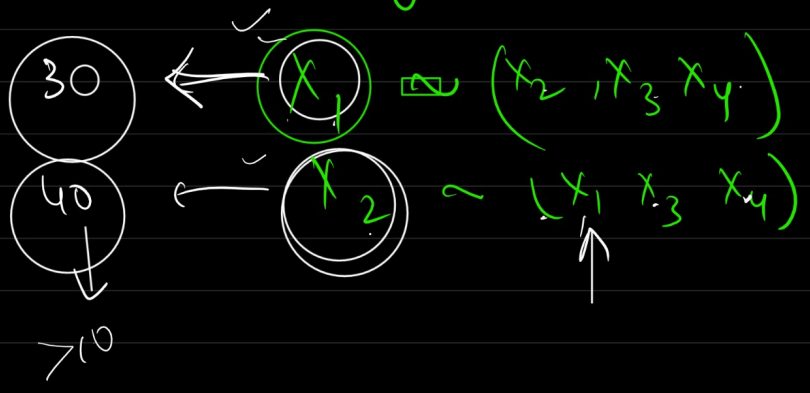
$X_4 \sim (X_1 X_2 X_3)$ — 2

VIF ~ 0 to ∞

VIF >= 10, then drop features one by one (starting with Highest Value)

① Why VIF >= 10 ?

② drop feature one by one ?



dropping one feature can change VIF of another feature

VIF > 10

$X_1 \sim (X_2 X_3 X_4)$

90% of variance in X_1 is explained by $X_2 X_3 X_4$

$$10 = \frac{1}{1-R^2}$$

$$1-R^2 = \frac{1}{10}$$

$$R_{sq} = 1 - \frac{1}{10} = \frac{10-1}{10} = \frac{9}{10} = 0.9$$

Feature VIF

X_1 13

X_2 12

X_3 8

X_4 7

< 10

> 10

First drop X_1 → Again calculate VIF ⇒

Kinds of multicollinearity

⇒ Data based multicol

⇒ Structural multicol

1000

1000 time VIF

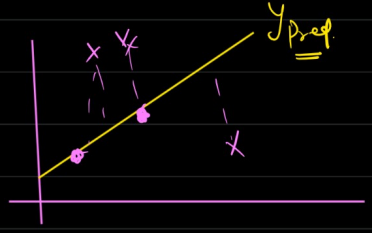
RFE — Recursive Feature Elimination

PCA

* Regularisation ⇒ To add something | to regularisation

To regularize

To penalize



data

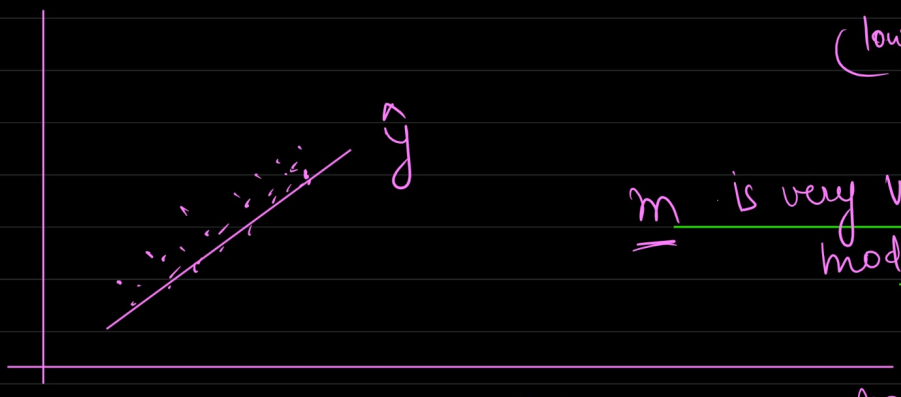
test

overfitting

train

Acc ↑ (low bias)

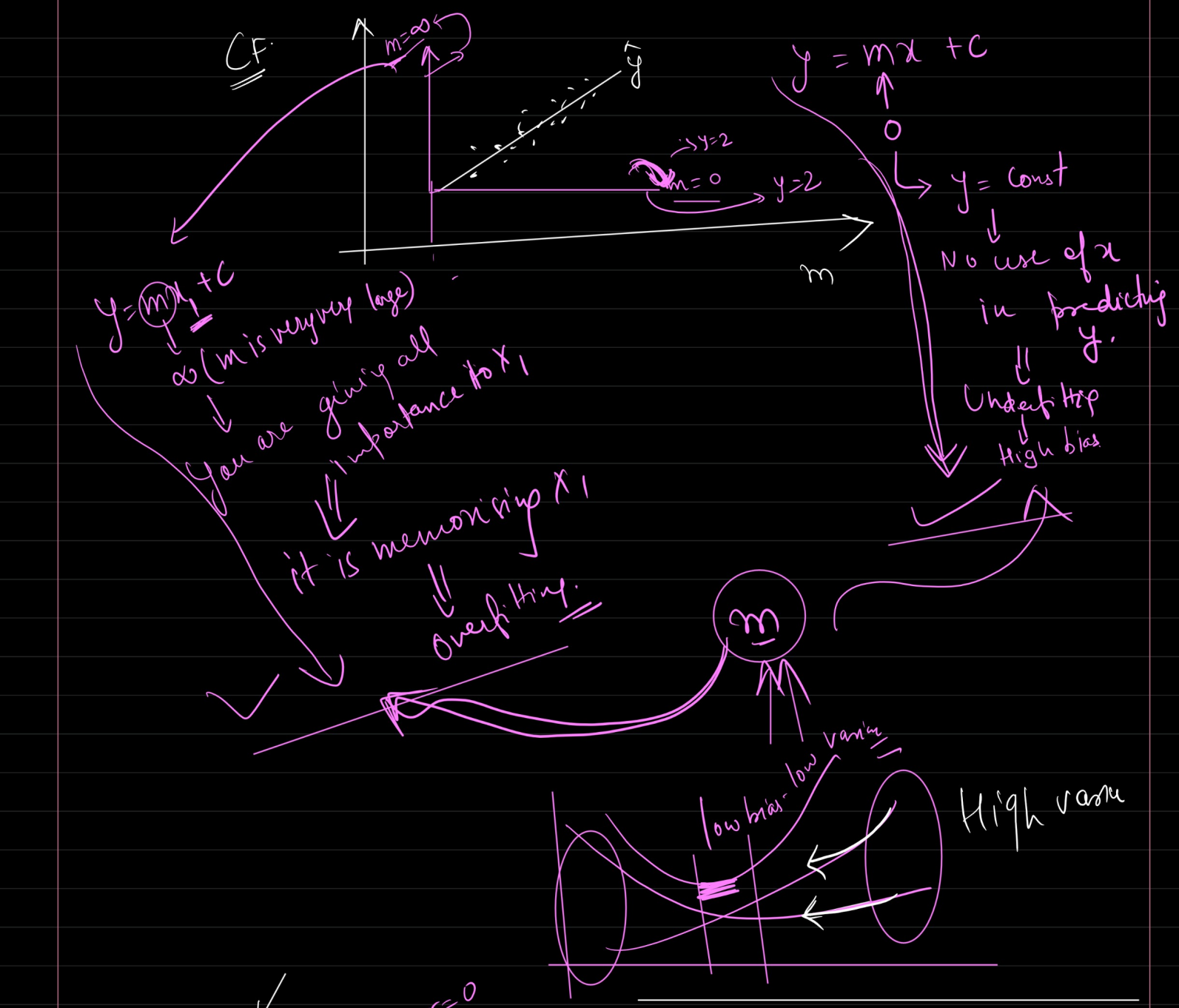
Acc ↓ (High variance)

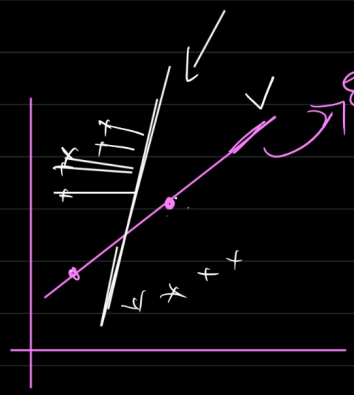


m is very very high → model has memorised the dp → it means overfitting

$y = 3.2 + 2.5x$

$y = 4.1 + 2.50x$





Overfitting

↓

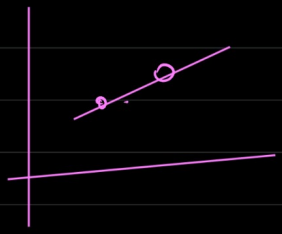
introduce some bias

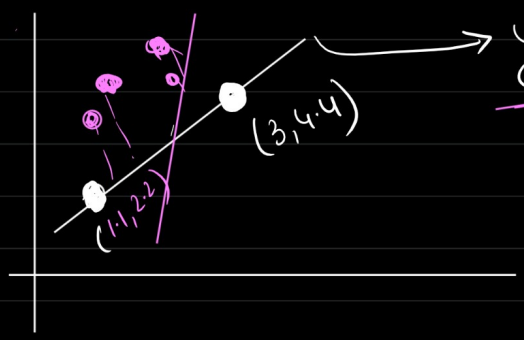
↓

error in the training.

✓ Cost fn = $\frac{1}{2n}\sum_{i=1}^{n}\left(y_{act} - (mx+c)\right)^2 + \lambda m^2$

Error = 0





$\rightarrow y = 1.1x + 2.1$

$CF = (y_{act} - y_{pred})^2$

$= 0$

CF >> high → line is not generic!

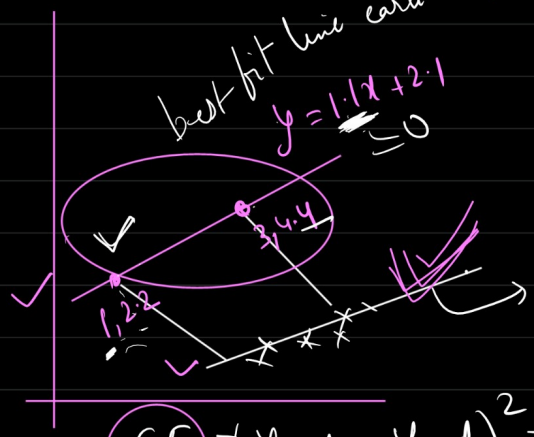
$$CF = \frac{1}{n}\sum_{i=1}^{n}(y_{act} - y_{pred})^2 + \hat{m}^2$$

$y_{pred} = mx + c$

best fit line earlier

$y = 1.1x + 2.1$

$= 0$



Hypothetical line $\hat{y} = 0.5x + 2$ (2.2)

Hypothetical line

$C.F = (2.2 - 0.5 \times 1 - 2)^2 + (4.4 - 0.5 \times 3 - 2)^2 + 1 \cdot (0.5)^2$

$CF = (y_{act} - y_{pred})^2 + \hat{m}^2$

best fit line: $\hat{m} = 1$

$0 + (1.1)^2$

$= 1.21$

$CF \Rightarrow (0.3)^2 + (0.9)^2 + (0.5)^2$

$= 1.15$

CF =

* Hypothetical line will be selected even if it is performing bad with train data because CF is lower as compared to other line.

multi → theory → Ridge and Elastic → Logistic

① Ridge — $CF + \lambda m^2$

② Lasso — $CF + \lambda |m|$

③ Elastic net — $CF + \lambda m^2 + \lambda |m|$