

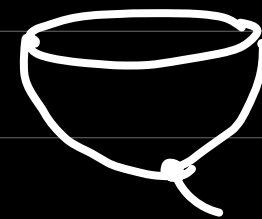
## Intro to ML and NN

### Lecture - Linear Regression - 5

Lin Reg - NO CHANCES OF GETTING STUCK IN LOCAL MINIMA

× local minima

learning rate



Global minima

$$w_j \rightarrow w_j - \alpha \frac{\partial L}{\partial w_j}$$

loss

$$\downarrow_m \quad \nearrow (y^i - \hat{y}^i)$$
$$\frac{-2}{m} \cdot \sum_i e_i \cdot x_j$$

$$b \rightarrow b - \alpha \frac{\partial L}{\partial b} \rightarrow \frac{-2}{m} \cdot \sum_{i=1}^m e_i$$

\*\*\*

$$\hat{y}_i = w_0 + w_1 x_1 + w_2 x_2 + \dots + w_d x_d$$

$$\downarrow$$

$$\frac{-2}{m} \cdot \epsilon$$

$$\propto$$

$$\downarrow$$

$$\frac{-2}{m} \cdot \epsilon \cdot x$$

$$\propto$$

$$\epsilon = \sum_{i=1}^m e_i$$

statsmodel library - significance of coefficients  
 - p-values  
 - Feature significant -  $p < 0.05$

Assumptions of Linear Regression (Statistics)

1. Assumption of linearity  $X \xrightarrow{f} y$   $f$ : linear
2. NO multi-collinearity

$f_1, f_2, f_3, f_4$

$$f_3 = \alpha + \alpha_2 f_1 + \alpha_3 f_2 + \alpha_4 f_4$$

linear combination of other

PREDICTORS

~~y~~  
Other  $x_j$

Solution: Remove some of the multi-collinear



VIF

---

# Variance Inflation Factor (VIF)

~~X~~  $\in \mathbb{R}^d$  - d features

~~y~~ original target  
not involved

$$f_1, f_2, f_3, \dots, f_d$$

Iteratively select each of the feature as **target**

e.g., step 1:  $f_1, f_2, f_3, \dots, f_{a-1}, f_a$

↓  
X

y↓

$$w_0 + w_1 f_1 + w_2 f_2 + w_3 f_3 + \dots + w_{d-1} f_{d-1} = f_d$$

 $R^2$ 

↓

Good  $\rightarrow f_d$  can be defined as L.C of  $\langle f_1, f_2, \dots, f_{d-1} \rangle$

$\Rightarrow$  YES Multicollinearity

$$VIF_j = \frac{1}{1 - R_j^2}$$

Target variable is  $j$ th feature.

Step 1) Calculate VIF for every feature

- $VIF_1$  - Target is  $f_1$
- $VIF_2$  - Target is  $f_2$
- $VIF_3$  - Target is  $f_3$
- ∴  $VIF_d$  - Target is  $f_d$

} Training d  
linear Reg  
Models.

Step 2) Remove one feature with highest VIF.

Case 1

$$R^2_j = 1$$

$$VIF = \frac{1}{1 - \cancel{R^2}} = \frac{1}{0}$$

$$VIF \rightarrow \infty$$

$\Rightarrow VIF \rightarrow \infty$  if strong multi-collinearity.

Case 2

$$R^2_j = 0$$

$$VIF = \frac{1}{1 - \cancel{R^2}} = \frac{1}{0}$$

$$= 1$$

$\Rightarrow VIF \rightarrow 1$  almost no multi-collinearity

CAUTION: Delete one feature at a time.

Delete the one with highest VIF (one of them)

Step 3) Repeat this process UNTIL WHEN???

Thumb Rule (multicollinearity)

1.  $VIF > 10$  V. strong MC, drop with high VIF
  2.  $5 \leq VIF < 10$  somewhat strong, drop (maybe)
  3.  $VIF < 5$  low MC, OK, Do not drop.
- 

Book Rules.

BUT, Removing certain feature is reducing performance  $R^2$  - Selling Price.

Check Performance of model again and see if

it has significantly dropped.

Case 1

Remove all  $t$  with  
 $VIF \geq 5$   
and performed  
didn't reduce  
A LOT

Case 2

After removing certain  
variable performance  
suddenly dropped

Stop at whichever happens first.

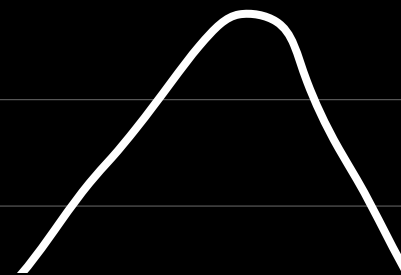
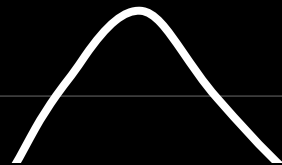
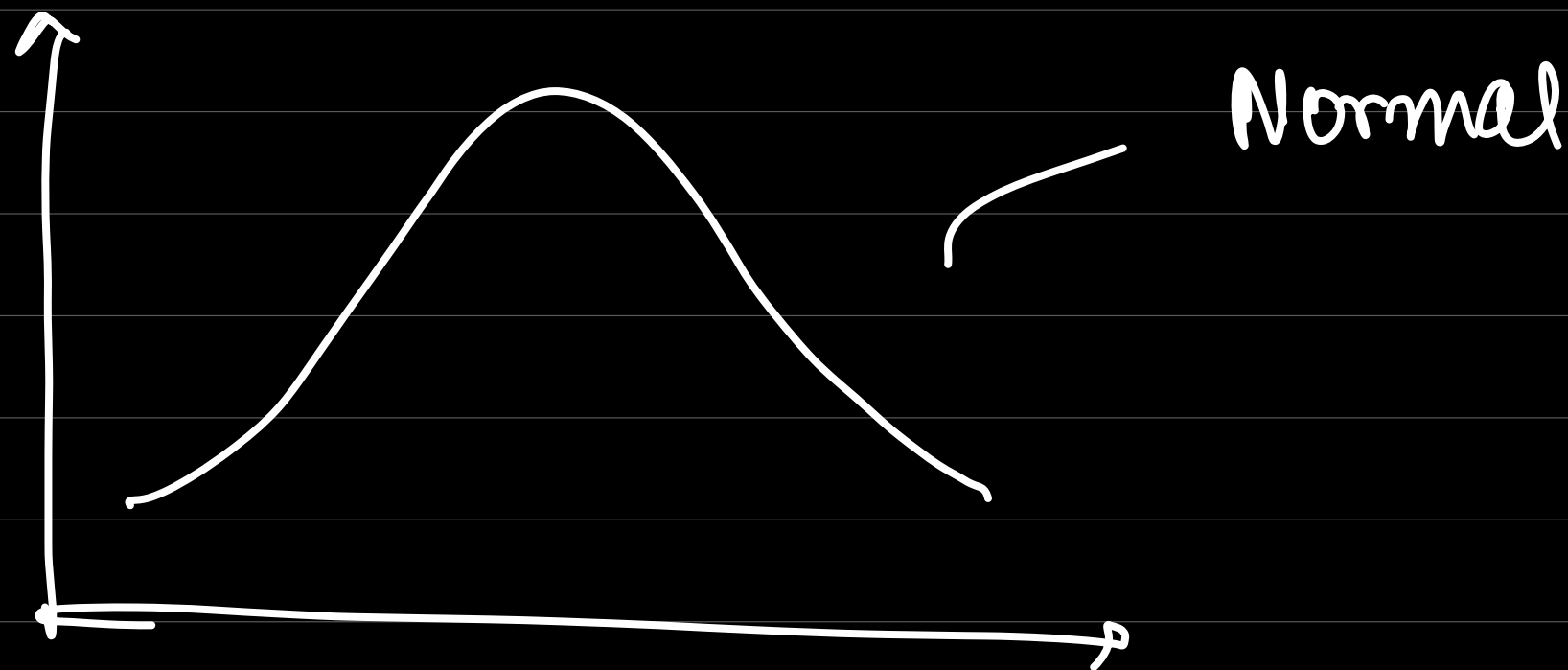


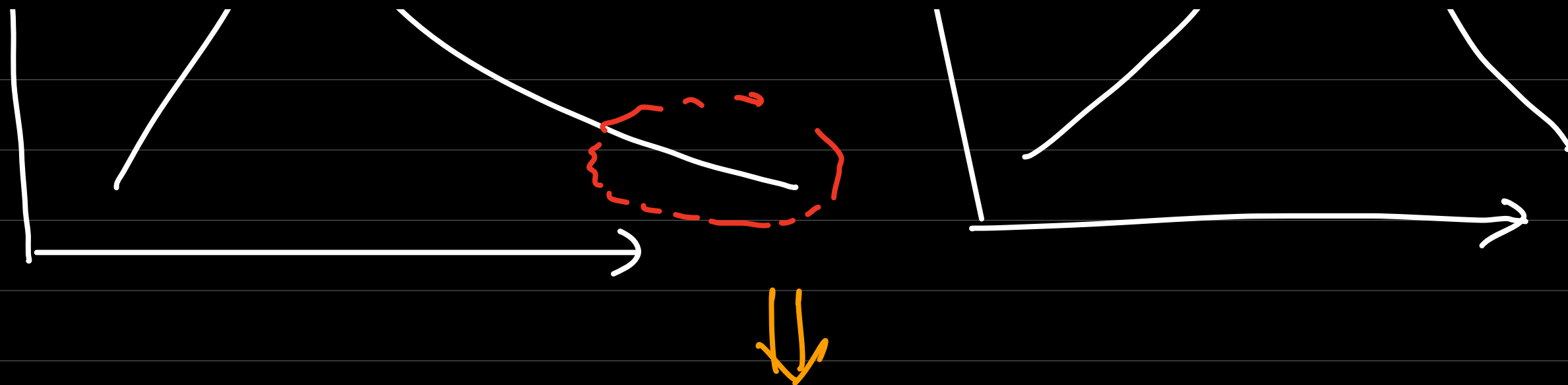
# # Assumption-3

error,  $e$ ,  $y^i - \hat{y}^i$

Normality of Residuals

$m$  examples  $\rightarrow e_1 e_2 e_3 \dots \dots e_m$





Error are not normal

↓ Signalling??

POTENTIAL PRESENCE OF OUTLIERS

For examples with high errors, drop them

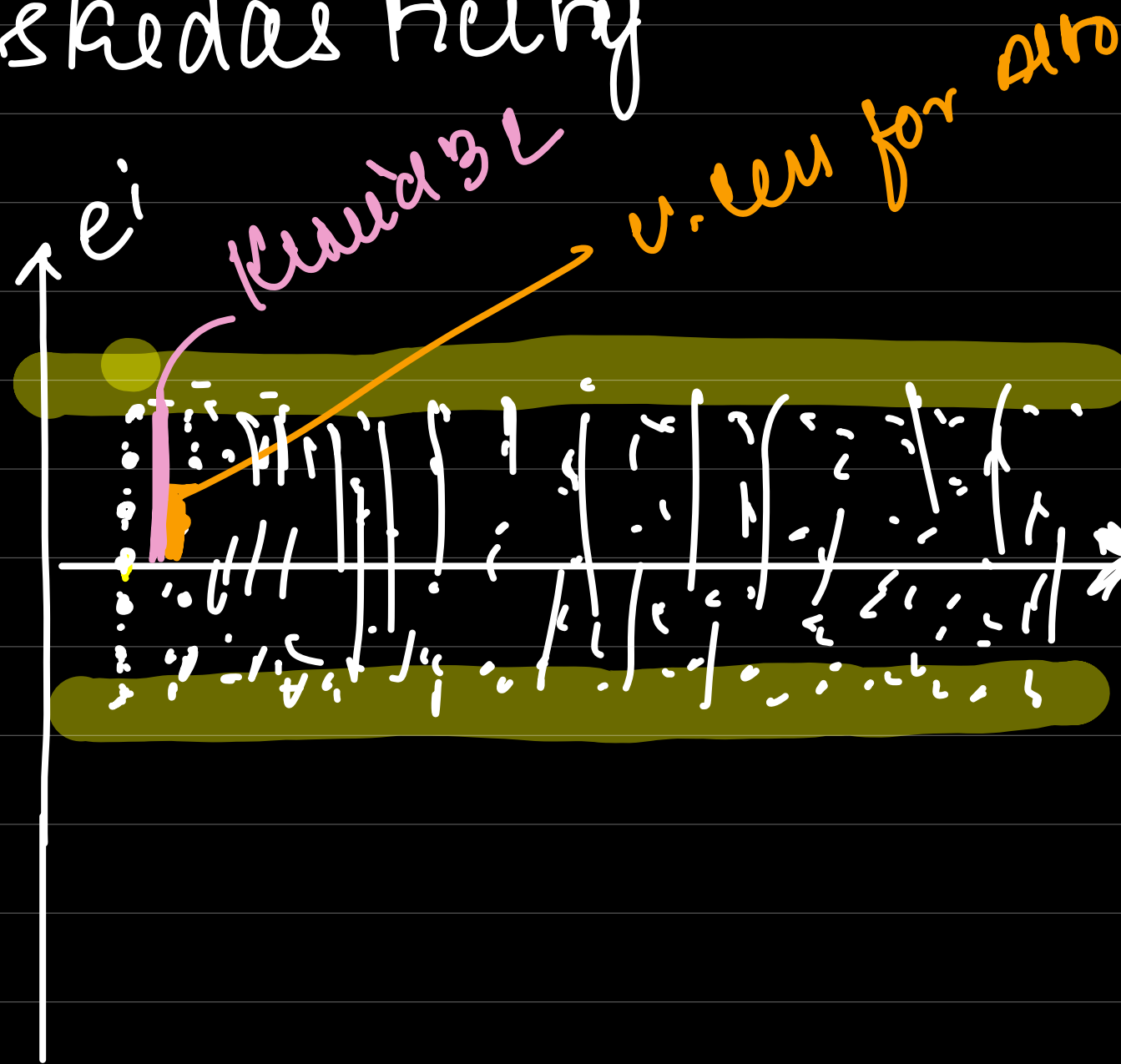
Re-train model.

Use: Shapiro Wilk Test.

# # Assumption 4

★★★★★

Heteroskedasticity should not exist  
Homoskedasticity



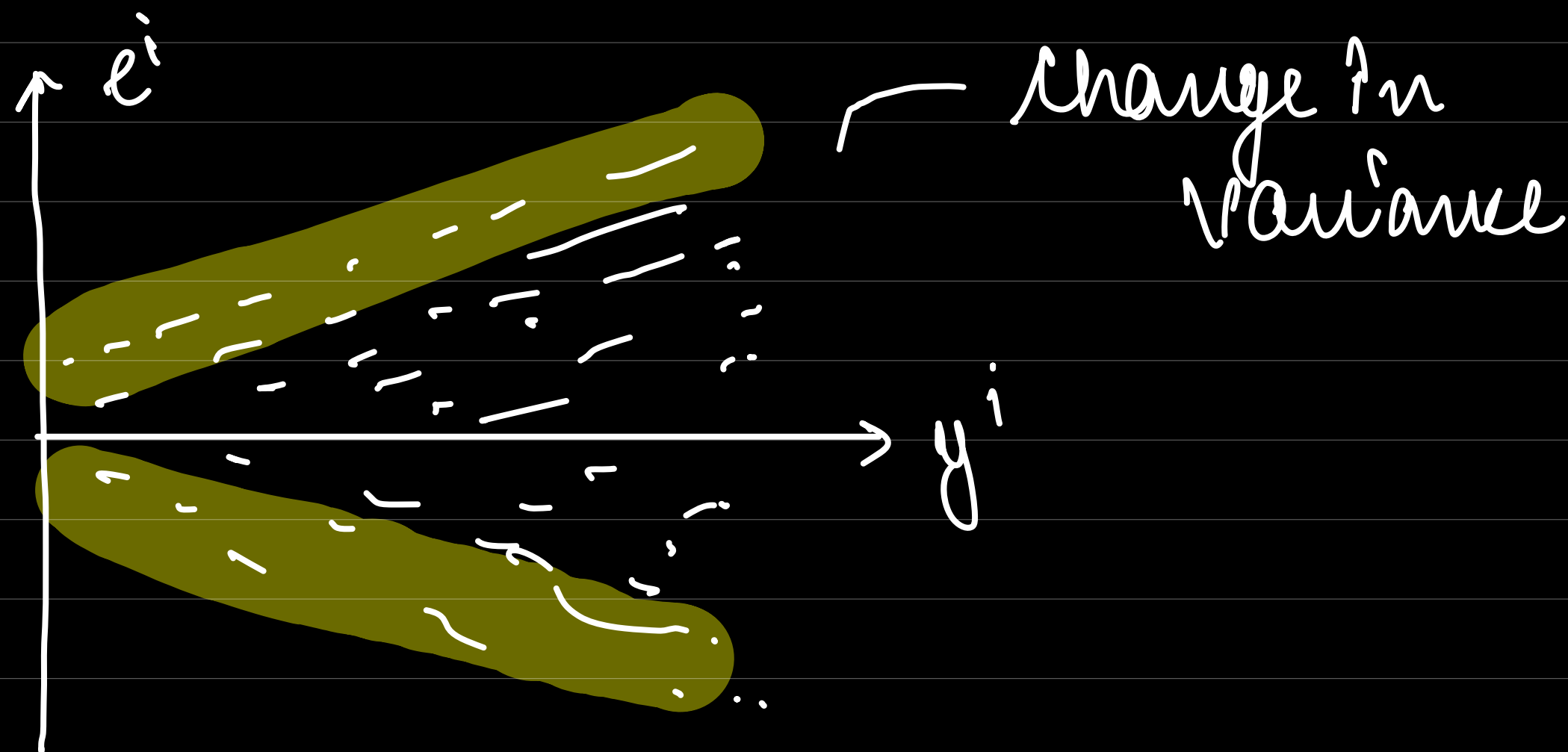
HOMOGENEITY

$y_i$  Target variable  
Selling Price

3L A110  
3L Kwid

1.0000 0 1 2 3 4 5 6 7 8 9 10

more increase in  $y(EP)$ , change the distribution of error



HETEROSDACITY

This should not happen.

Plot and analytical check. - Goldfedd Quant

# Assumption-5

No autocorrelation.

??

Time-component

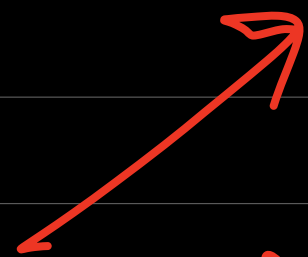
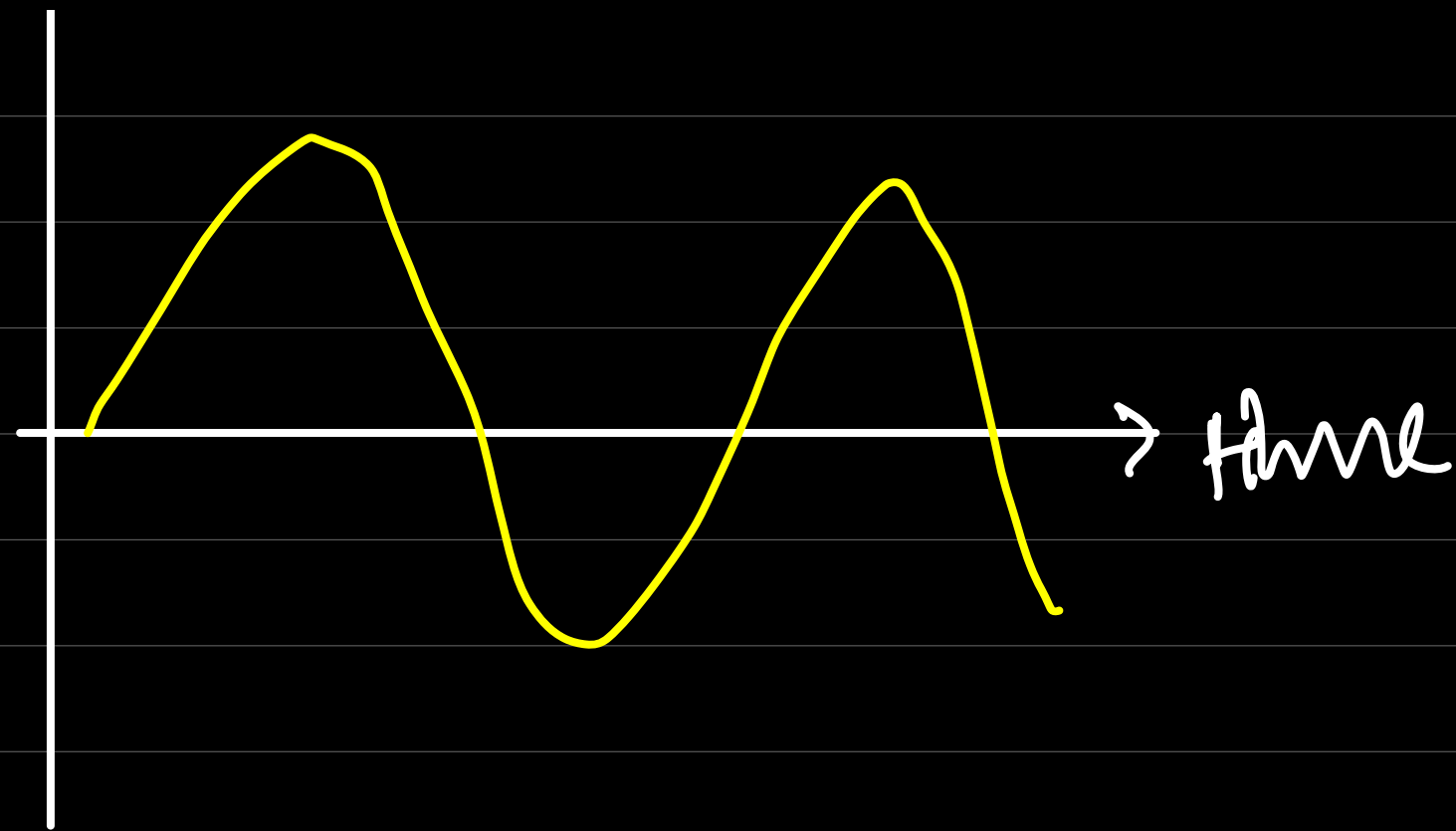
Time-series data.

$n$ -samples.

$x_1 \ x_2 \ x_3 \ x_4 \dots x_n$

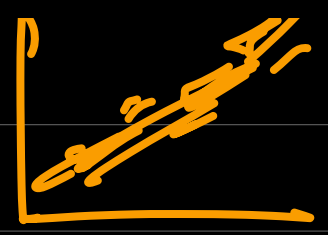
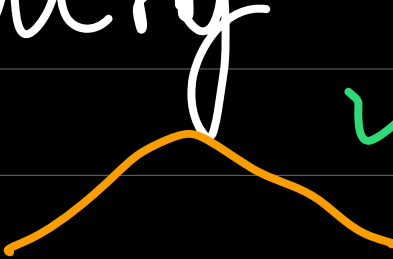
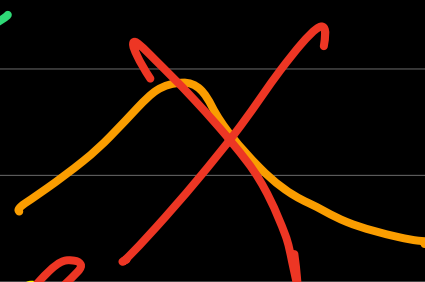
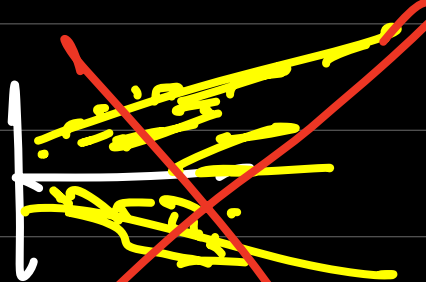
$t_0 \ t_1 \ t_2 \ t_3 \dots$

error  
↑



No pattern should be  
visible here.

Summary

- 1) linear function  $X \rightarrow y$  
- 2) NO Multi-collinearity
- 3) Errors should be  
- 4) NO Heteroskedasticity 
- 5) NO Auto-correlation 