Importance of error/residuals plots

01 August 2025 15:18

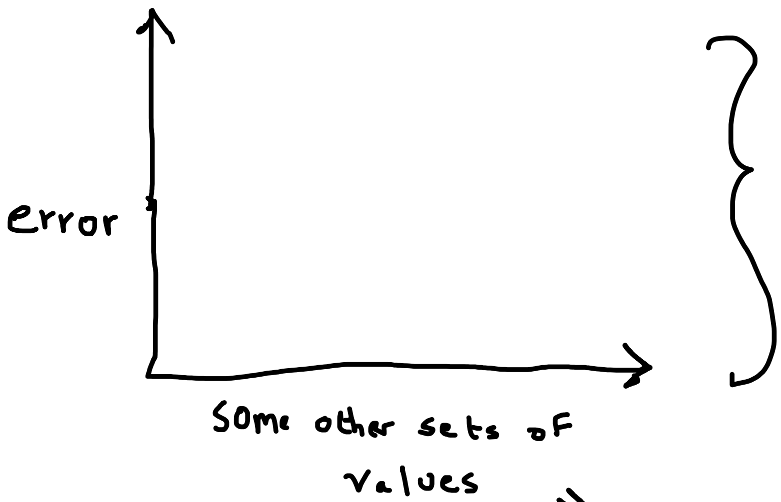
	y_actual	y_pred	error
0	50.000000	50.252809	0.252809
1	51.010101	53.509858	2.499757
2	52.020202	47.040657	-4.979545
3	53.030303	56.498296	3.467993
4	54.040404	51.948896	-2.091508

→ y pred - y actual

⇓

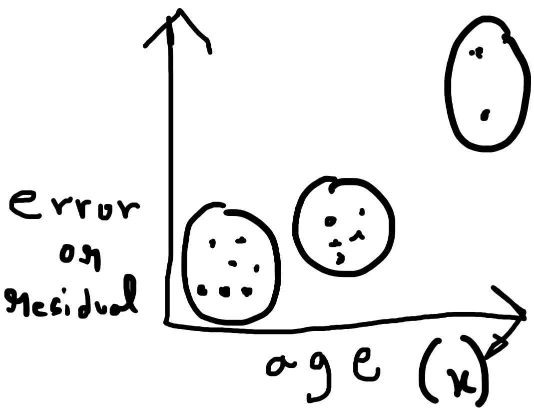
residual

the importance of error based plots !



⇒ Can be input variables or target variable

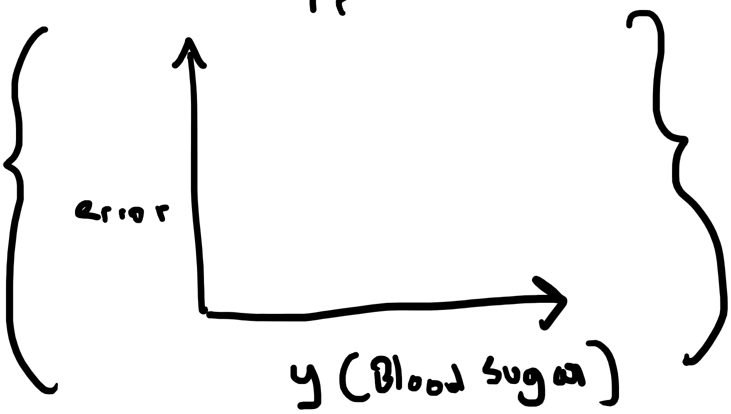
Assume age is a feature



⇒ compare errors at different of x variable

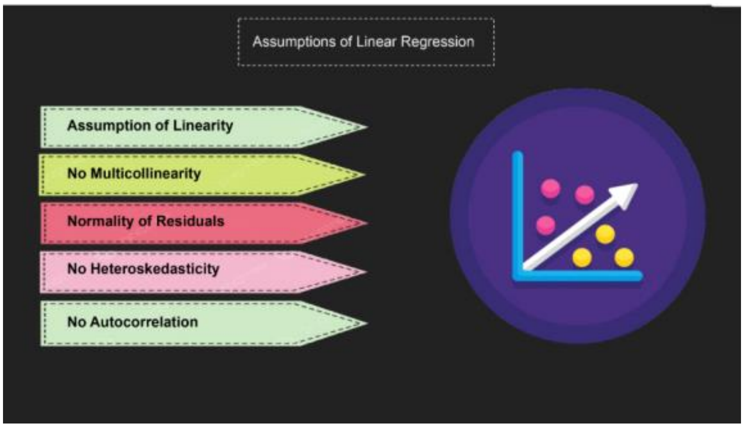
Throughout the session, we will look into importance of error vs y variable plots

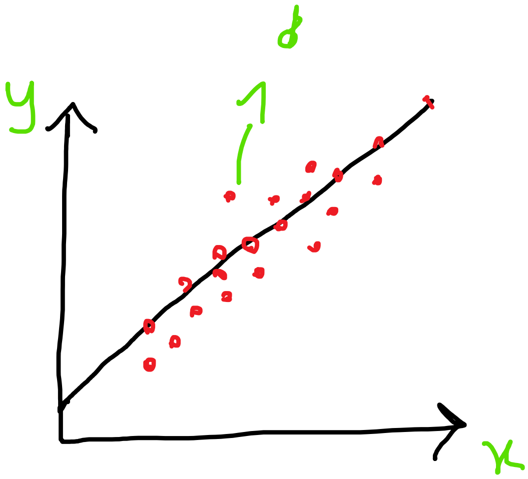
⇑

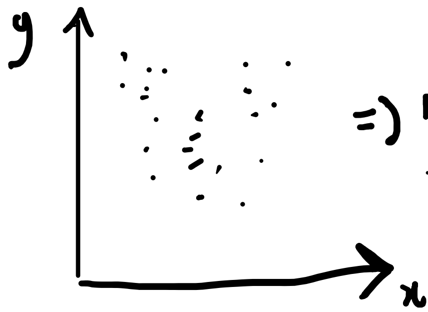


Assumption of Linearity

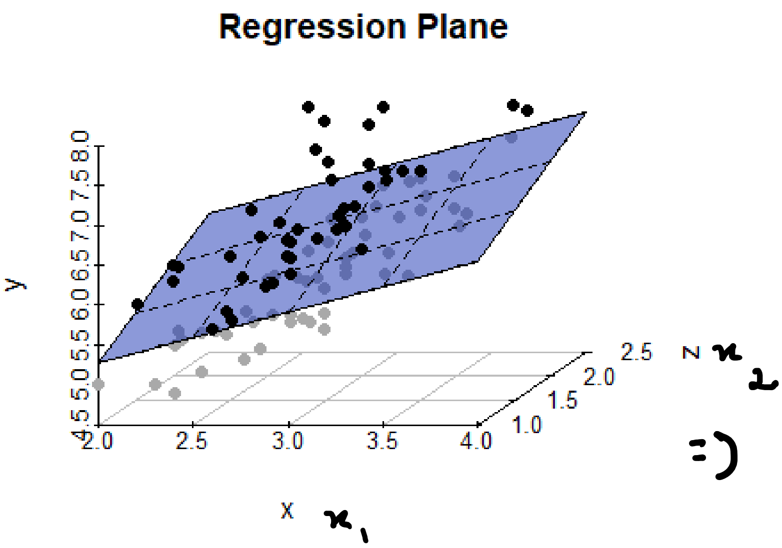
30 July 2025 16:02



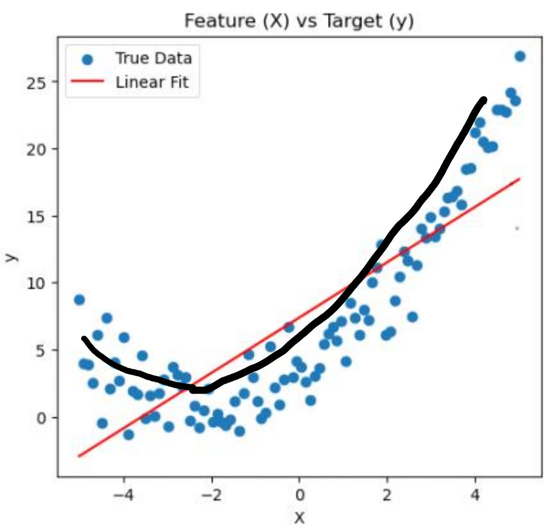


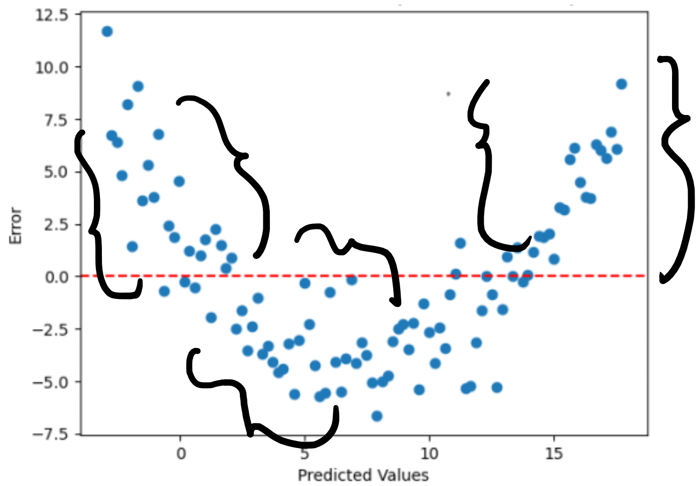


=) Not linear



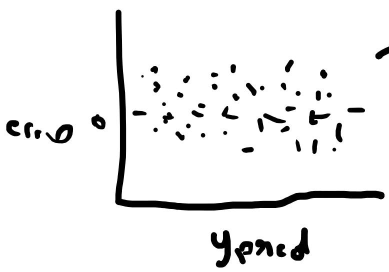
=) hyperplane also linear





When there is a noticeable relationship b/w ypred and error, linear regression is not suitable option

ideal Scenario



→ random plot } → linear regression working well

The trade off at times

01 August 2025 16:42

Statistical Correctness

vs

Predictive Correctness

Sometimes a trade-off with linear models

No Multi-collinearity

31 July 2025 22:58

What is Multi-collinearity?

Travel_expense $= \omega_1 \times Km + \omega_2 \times Fuel + \omega_0$

Features | | target

Km_Driven (X1)	Fuel_Consumed (X2)	Travel_Expense (Y)
1000	80	6400
800	64	5120
1200	96	7680
600	48	3840
900	72	5760

α

→ Are X_1 and X_2 related?

$$\{ x_1 = w x_2 + b \} \rightarrow \text{linearly related}$$

Why is it an issue?

→ Feature interpretability heavily impacted, Which is more important, x_1 or x_2? Both?

→ Sign flipping in co-efficients - Causes Confusion.

But it is not always a problem!

How can we avoid Multi-collinearity?

VIF! → Variance Inflation Factor

How to deal with multicollinearity?

We will use **Variance Inflation Factor (VIF)**

Say, we have 'd' features $< f_1, f_2, f_3 \ldots\ldots\ldots f_d >$

In, **(VIF)** we treat
- one feature as 'y'
- remaining features as 'x'

$f_1, f_2, f_3 \ldots\ldots\ldots f_d$	f_4
x_i	y

F_1, F_2, F_4 | F_3

F_3, F_1, F_4 | F_2

{ identify Multi-collinearity }

$x_1, x_2, x_3, x_4 \ldots\ y$

Now,

Train **linear regression** model with (x_i, y)

Find R^2 of the model

To Calculate VIF :

$$VIF = \frac{1}{1 - R_j^2},\ R_j^2 \ : R^2 \text{ for } j^{th} \text{ feature}$$

Features

target

$$x_1 = w_1x_2 + w_2x_3 + w_3x_4 + w_0 \Rightarrow$$

$\frac{1}{1 - R\,sq_{x_1}}$ For this model = VIF for x_1

$$x_2 = w_1x_1 + w_2x_3 + w_3x_4 + w_0$$

⋮

$$x_d = w_1x_1 + w_2x_2 \ldots\ldots w_0$$

$1 \rightarrow R^2 = 0.9$

$$\left\{ VIF = \frac{1}{1 - 0.9} = \frac{1}{0.1} = 10 \right\}$$

Higher the VIF For a Variable, higher the Multi-collinearity it shares with other variables.

VIF > 5 ⇒ think of dropping

VIF > 10 ⇒ {definitely drop}

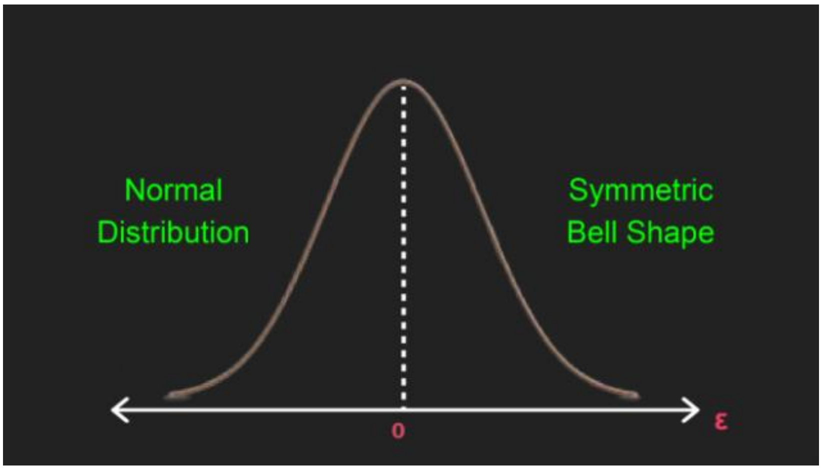
{ 0 to ∞ }

Errors are normally distributed

31 July 2025 22:58

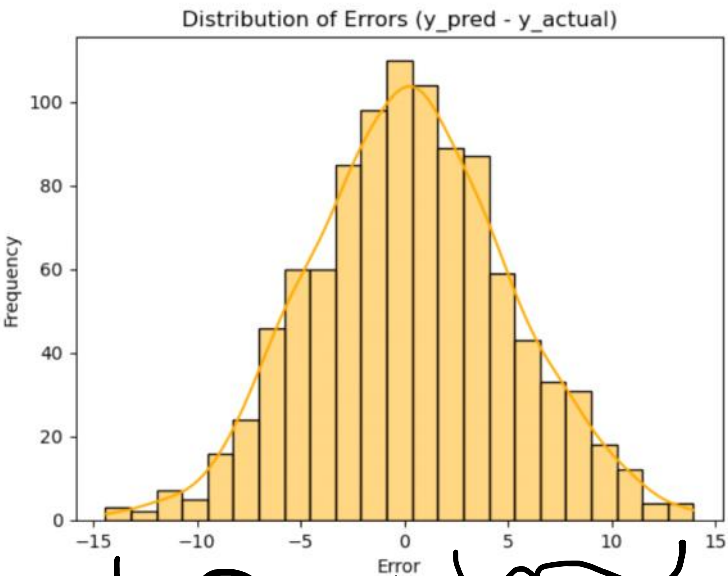
$$\text{error} = \underset{\text{Actual}}{y_i} - \underset{\text{Predicted}}{\hat{y}_i}$$

Frequency



Why ?

i)



Symmetric

$\Downarrow$

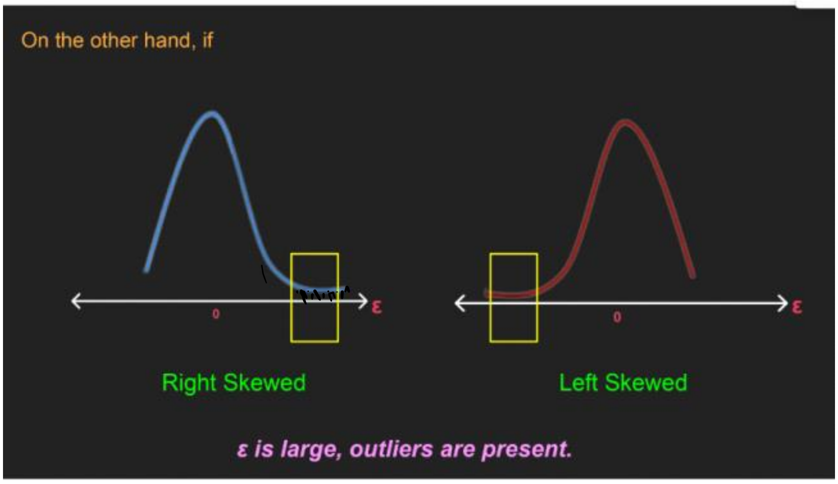
No. of overpredicted = No. of underpredicted

$\underbrace{\hphantom{xxxxxxxx}}_{\text{Underprediction}}$ $\underbrace{\hphantom{xxxxxxxx}}_{\text{overprediction}}$

{ypred - yactual}

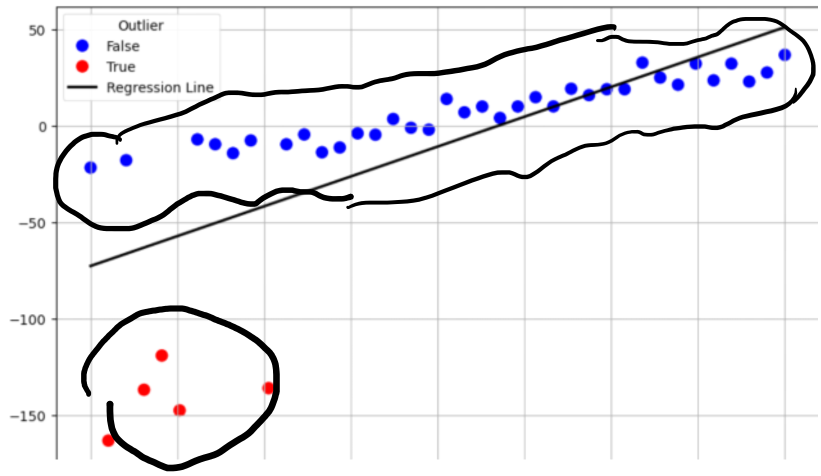
If my model is either o.p or u.p, heavily

On the other hand, if



i) Model is unable to capture sufficient information

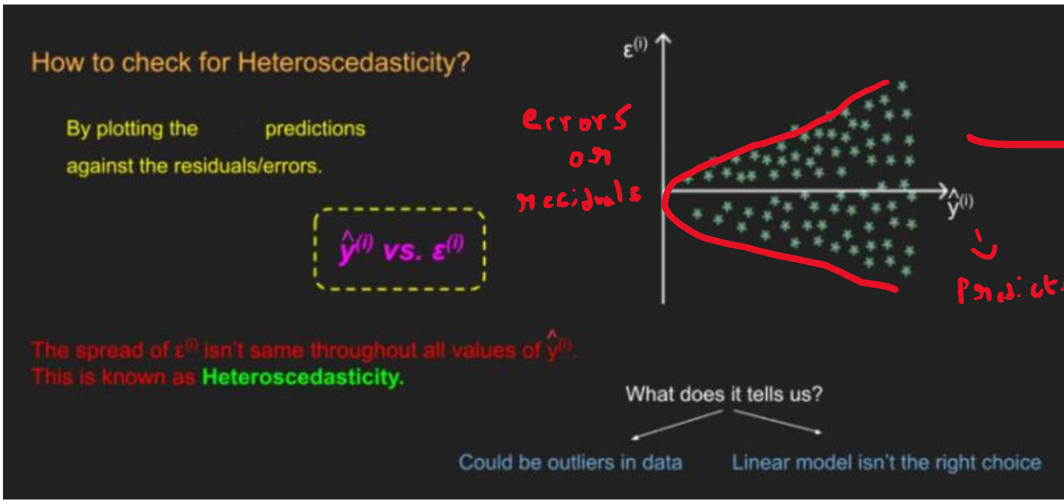
ii) There are outliers in the data



No Heteroscedasticity

01 August 2025 16:09

Variance of errors should be constant across y-predicted



Predicted

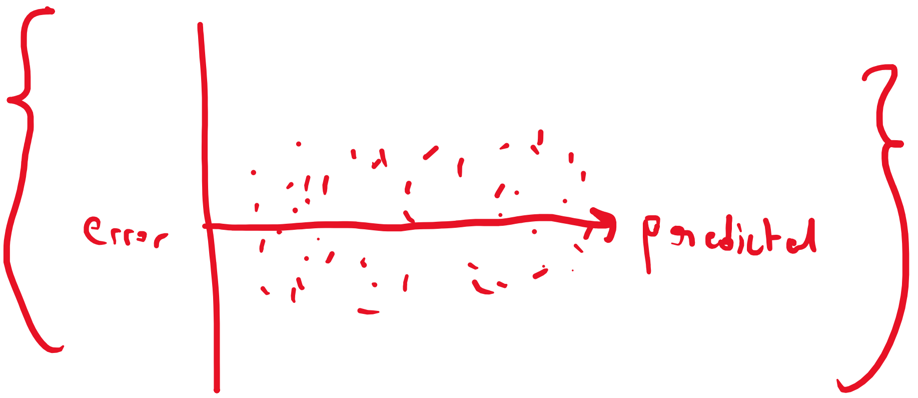
here it is not constant

What does Heteroscedasticity indicate?

i) Linear Model may not be suitable

ii) Use a non-linear Model

iii) Introduce other variables into your model



→ Homoscedasticity

No Autocorrelation of errors

01 August 2025 16:12

What is auto - correlation ?
self

in errors there should be no auto-correlation

Note : Assumption applicable only For time-series data

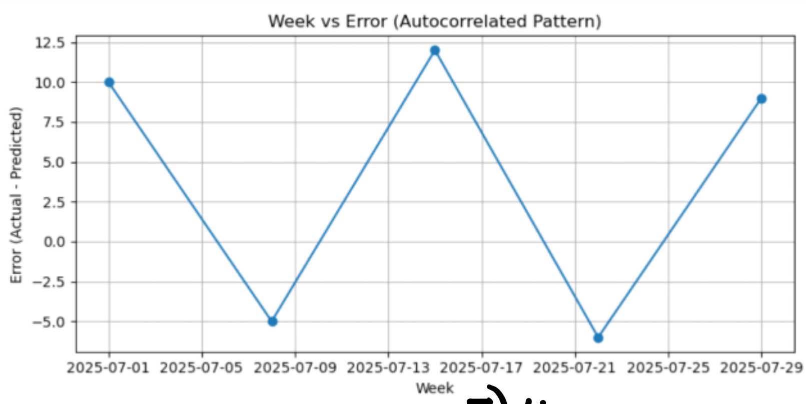
What is time-series data ?

Week	Sales (target)	Promotion	Temperature (°C)	Holiday
2025-07-01	200	1	30	0
2025-07-08	180	0	31	0
2025-07-15	220	1	29	0
2025-07-22	250	1	28	1
2025-07-29	190	0	32	0

} Predicting based on a time component is called time-series

📊 Final Table

Week	Sales (Actual y)	Predicted Sales (y_pred)	Error (y - y_pred)
2025-07-01	200	190	10
2025-07-08	180	185	-5
2025-07-15	220	208	12
2025-07-22	250	256	-6
2025-07-29	190	181	9



→ time component

Error
vs
time component

odd weeks errors are correlated

even week errors are correlated

How is this similar to heteroscedasticity?

trend in errors!!