Linear Regression - 04

Assumptions of Linear Regression

→ Assumption of Linearity

→ No Multi-Colinearity

→ Normality of Residuals $(y-\hat{y})$

→ No Heteroskadasticity

→ No Autocorrelation

No Multicolinearity

Colinearity ?

f_1 , f_2

if $\boxed{f_2 = \alpha f_1 + \beta}$

f_1 & f_2 are colinear

age , year

age = 2024 - year

Colinearity multiple features = Multi-colinearity

f_1	f_2	f_3	f_4

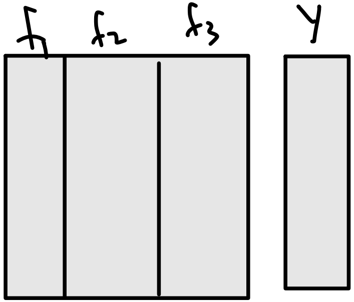
$$f_2 = \alpha_1 \cdot f_1 + \alpha_3 \cdot f_3 + \alpha_4 \cdot f_4 + \alpha_0$$

f_2 is multicolinearity

Height (cms)	Height (ft.)

No Multicolinearity !!

Q → why MC is a Problem?



M.C exists if $\boxed{x_2 = 1.5 x_1}$

Train | Optimized → $w^* = [w_1, w_2, w_3], w_0$

$w^* = [1, 2, 3], w_0 = 5$

$x_q = [x_1, x_2, x_3]$

① : $\hat{y} = 1.x_1 + 2.x_2 + 3.x_3 + 5$ → $w = [1, 2, 3]$

↓

$\hat{y} = 1.x_1 + 2.(1.5x_1) + 3.x_3 + 5$

$$\hat{y} = 1.x_1 + 3.x_1 + 3x_3 + 5$$

② : $\hat{y} = 4x_1 + 3x_3 + 5$ → $w = [4, 0, 3]$

$x_q = [2, 3, 1]$

① $1\times2 + 2\times3 + 3\times1$

$\hat{y} = 2 + 6 + 3$

$\hat{y} = 11$

② $\hat{y} = 4\times2 + 3\times0 + 3\times1$

$\hat{y} = 8 + 0 + 3$

$\hat{y} = 11$

Messed up with weights → Unstability of weights

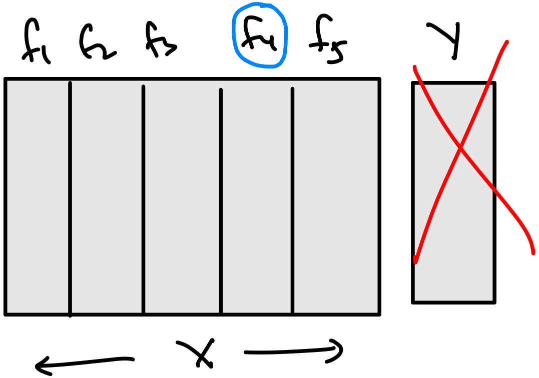
→ No feature importances

→ No interpretability

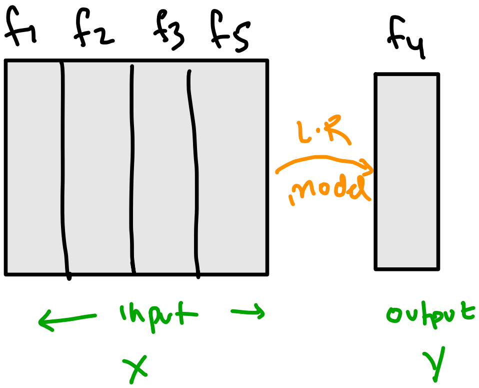
VIF [Variance Inflation factor]

$\boxed{f_j}$

$\leftarrow w_j \rightarrow$



VIF(f4)



$$\boxed{\hat{f}_4 = w_1 f_1 + w_2 f_2 + w_3 f_3 + w_5 f_5 + w_0}$$

R2 Score

if R2 = 0.98 : M.C exists

if R2 = 0.18 : m.c Not exists

$$\boxed{VIF_j = \frac{1}{1 - R^2_j}}$$

if $R_2 \to 1$ $\quad VIF \to \infty$

$R_2 \to 0$ $\quad VIF \to 1$

$$VIF \to [\underset{\text{No. M.C}}{1}, \underset{\text{v. large M.C}}{\infty})$$

THUMB RULE :

$VIF > 10$: v. High M.C

$VIF > 5$: High M.C

$VIF < 5$: low M.C

$VIF(f_1) =$ —

$VIF(f_2) =$ —

⋮

$VIF(f_d) =$ —

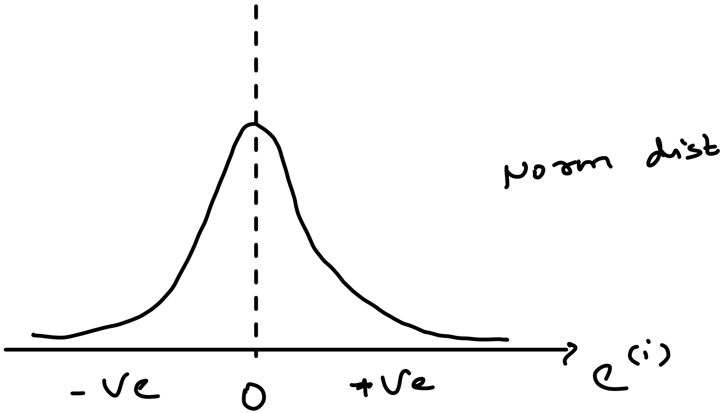
Independent features

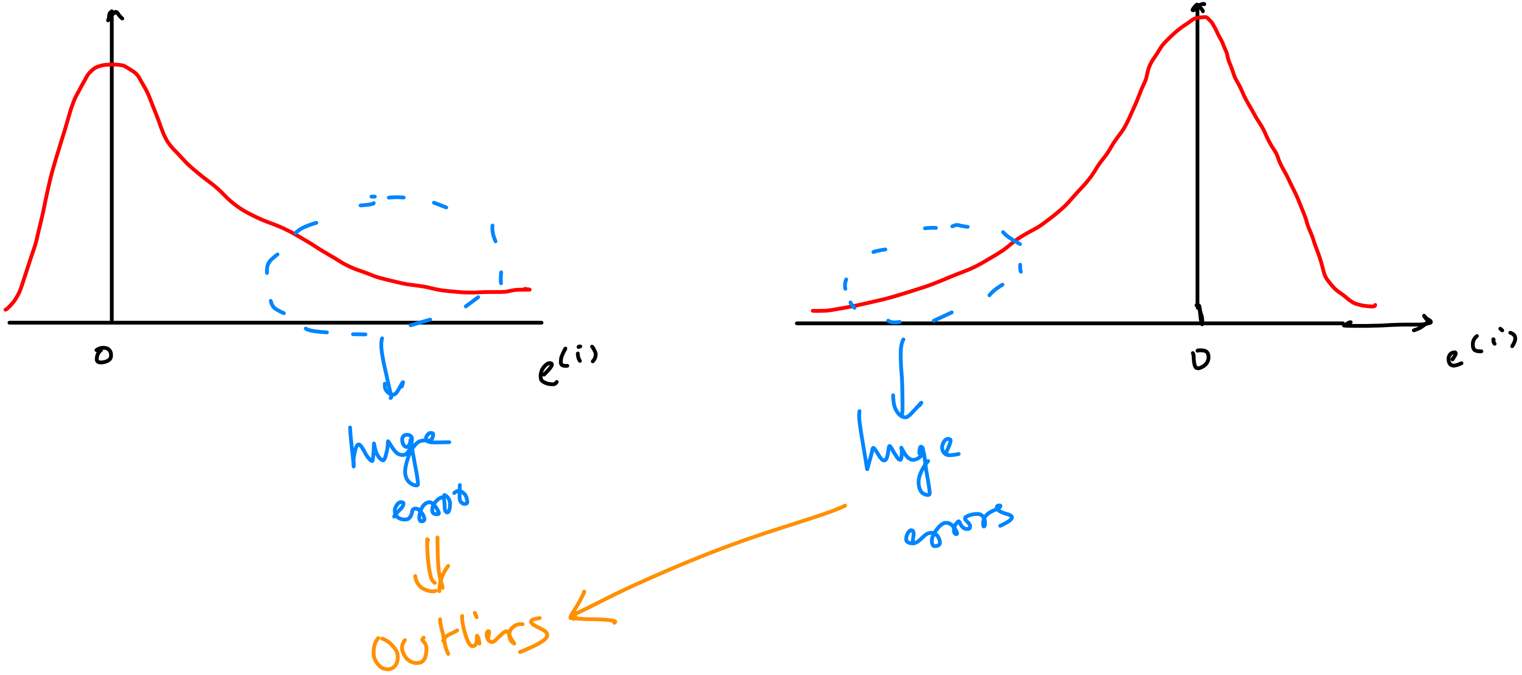
Normality of Residuals

$\to y - \hat{y}$

$$y = \underset{\hat{y}}{\underset{\downarrow}{\boxed{w^T x + w_0}}} + \epsilon$$

$$\boxed{\epsilon = y - \hat{y}}$$ residuals.



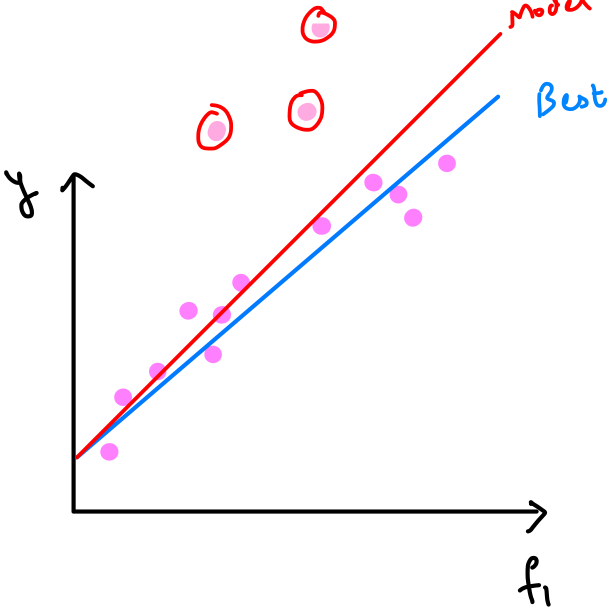


0
$e^{(i)}$
huge error
outliers
0
$e^{(i)}$
huge errors

Impact of Outliers

I. identify outliers ?

II. Deal outliers !

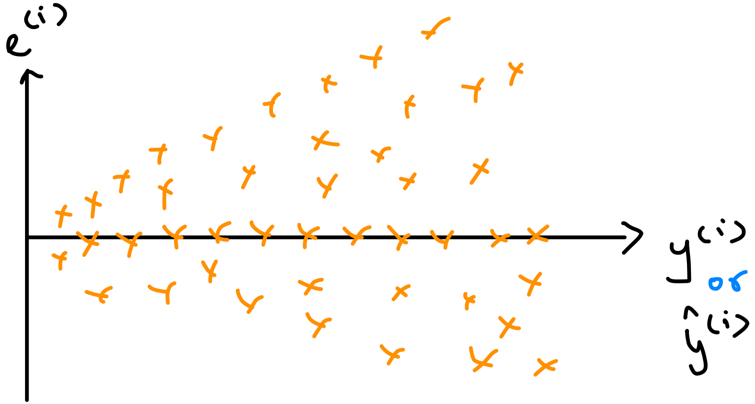


Residual Analysis

No Heteroskadasticity

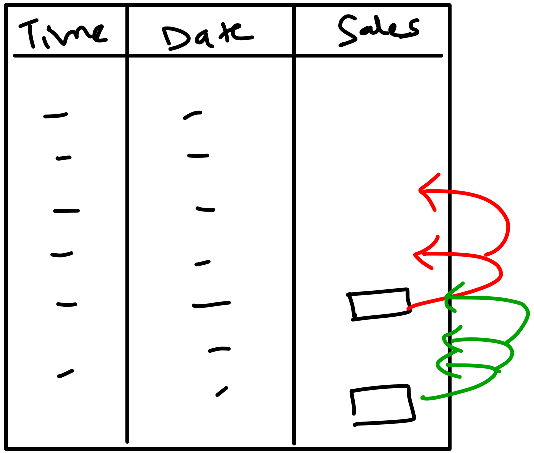
$\epsilon^{(i)}$ Vs. $y^{(i)}$

Homoskedasticity



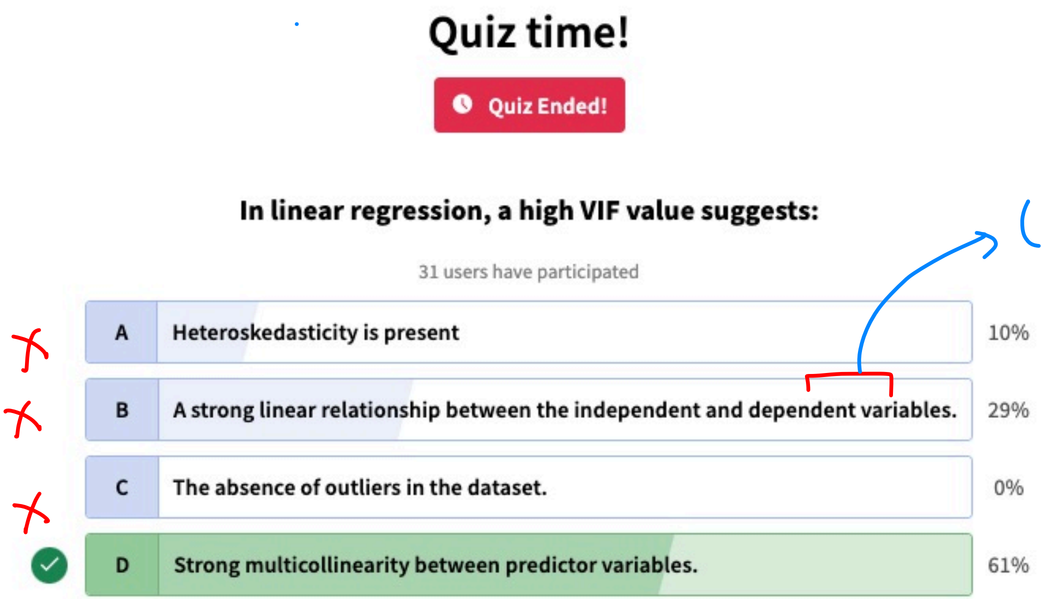
No Auto Correlation

→ in TimeSeries



Correlation with datapoints itself

→ every data point should be independent



Quiz time!

Quiz Ended!

In linear regression, a high VIF value suggests:

31 users have participated

✗	A	Heteroskedasticity is present	10%
✗	B	A strong linear relationship between the independent and dependent variables.	29%
✗	C	The absence of outliers in the dataset.	0%
✓	D	Strong multicollinearity between predictor variables.	61%

(y) e.g → Price