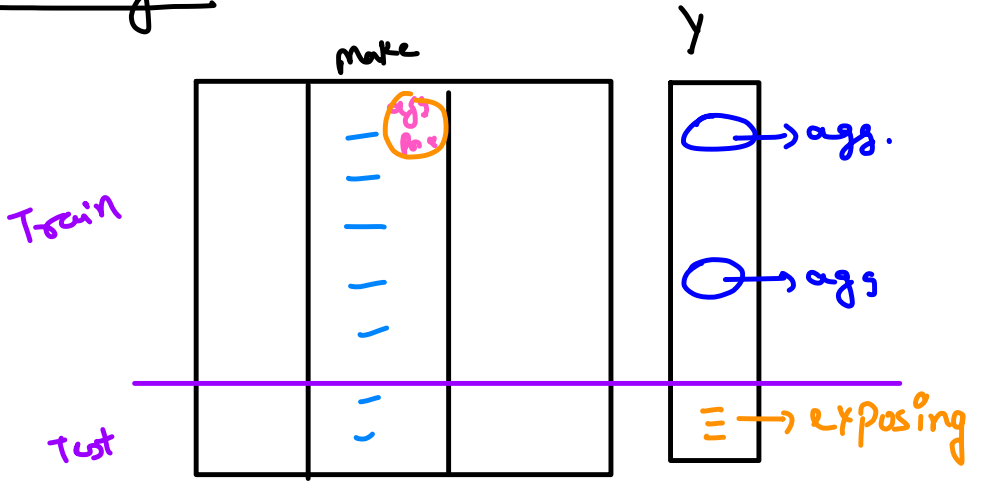
Linear Regression - 04

Data leakage



{ Maruti : 7.4
Honda : 8.1
TATA : 9.2 }

first Train / Test Split

Assumptions of Linear Regression

- Assumption of Linearity
- No Multi-Colinearity
- Normality of Residuals $(y - \hat{y})$
- No Heteroskadasticity
- No Autocorrelation

No Multicolinearity

Colinearity ?

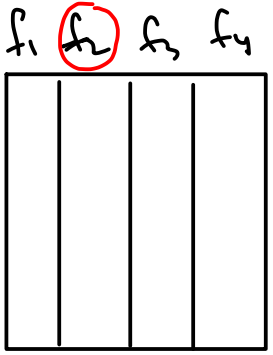
f_1, f_2 if $\boxed{f_2 = \alpha \cdot f_1 + \beta}$

f_1 & f_2 are Colinear.

age, year

$$age = -1 \cdot Year + 2025$$

Colinearity multiple feature ⇒ Multi-Colinearity



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

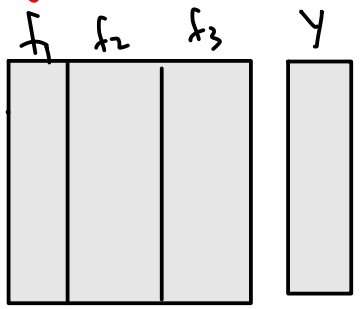
f_2 is multicolinearity

height (cm) | height (ft)

$$f_1 = \alpha_2 \cdot f_2 + \alpha_0$$

No multicolinearity !!

Q → Why MC is a Problem ?



if M.C exists

$f_2 = 1.5 x_1$

optimized → $w^* = [w_1, w_2, w_3], w_0$

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

$x_q \rightarrow [x_{q1}, x_{q2}, x_{q3}]$

$\hat{y}_q = w_*^T x_q + w_0$

① — $= 1 \cdot x_{q1} + 2 x_{q2} + 3 x_{q3} + 5$ $\quad w = [1, 2, 3]$

$= 1 \cdot x_{q1} + 2(1.5 x_{q1}) + 3 x_{q3} + 5$

(2) — = $4x_{q_1} + 3x_{q_3} + 5$ $\quad w = [4, 0, 3]$

$x_q = [2, 3, 1]$

(1) $1.2 + 2.3 + 3.1 + 5$

$\hat{y} = 2 + 6 + 3 + 5$

$= 16$

(2) $4 \cdot 2 + 3 \cdot 1 + 5$

$\hat{y} = 8 + 3 + 5$

$\hat{y} = 16$

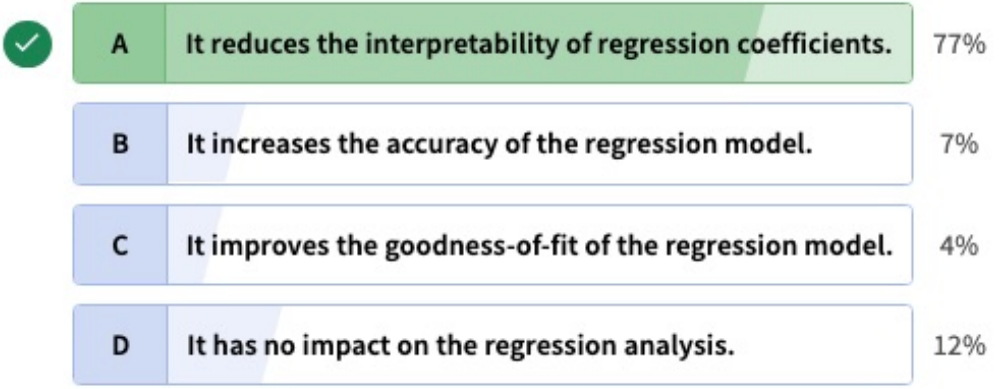
→ No. feature importance

→ interpretability

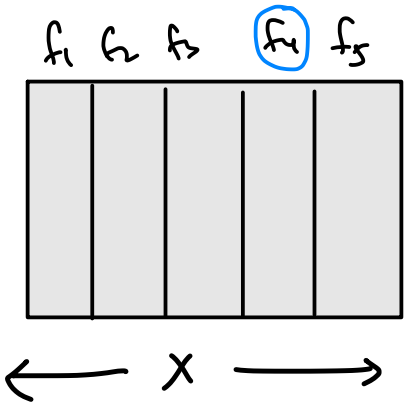
Messed up with weights → unstability of weights.

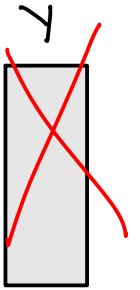
How does multicollinearity affect regression analysis ?

84 users have participated

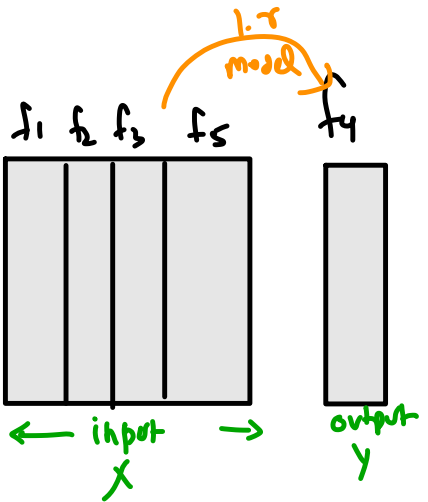


/// Variance Inflation factor [VIF]





VIF(f4)



$f_4 = w_1 \cdot f_1 + w_2 \cdot f_2 + w_3 \cdot f_3 + w_5 f_5 + w_0$

R2 Score

if $R_2 = 0.18$: M.C Not Exists

if $R_2 = 0.92$: M.C exists.

$$VIF_j = \frac{1}{1-R2_j}$$

if $R_2 = 0$; $VIF \to 1$

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

$VIF \in [1, \infty)$

- 1 → No.M.C
- ∞ → M.C exists. v.large

$VIF(f_1) =$ —

$VIF(f_2) =$ —

⋮

$VIF(f_d) =$ —

THUMB RULE :

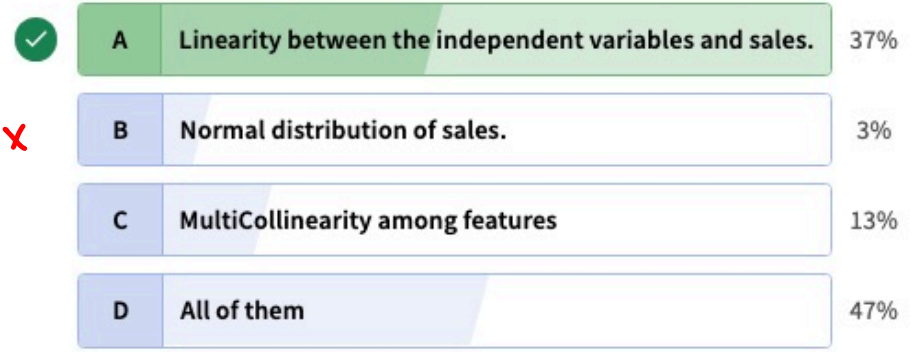
$VIF > 10$: v.high m.c

$VIF > 5$: m.c

$VIF < 5$: No worry

A clothing store wants to predict sales based on factors like price, promotions, and store location.
Which assumption of linear regression is important for accurate sales predictions?

87 users have participated



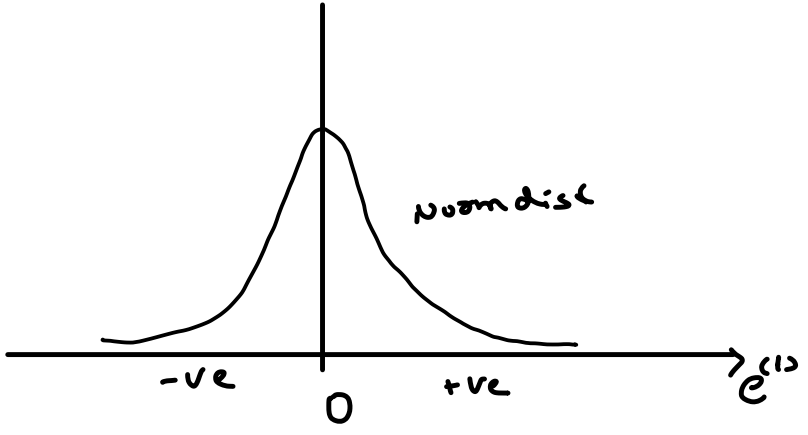
Normality of Residuals

→ $y - \hat{y}$

$$y = \omega^T x + \omega_0 + \epsilon$$

$$\hat{y}$$

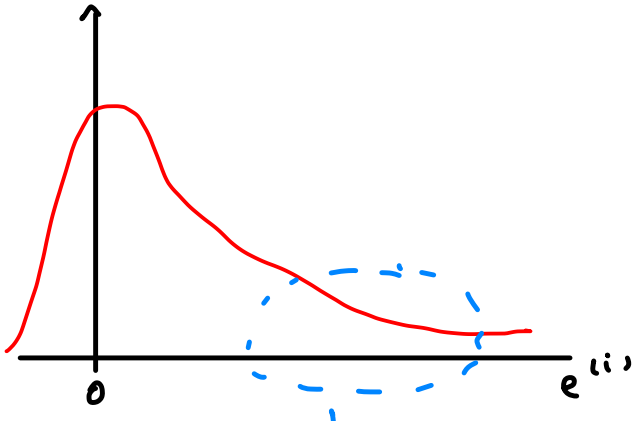
$$\epsilon = y - \hat{y}$$ residuals



3.5L - 4L

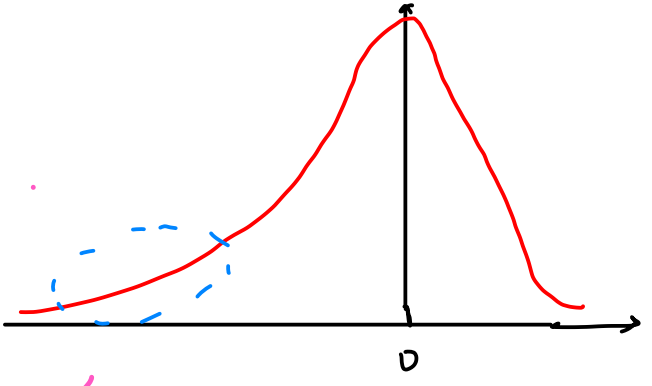
↓ → over predicted

-ve

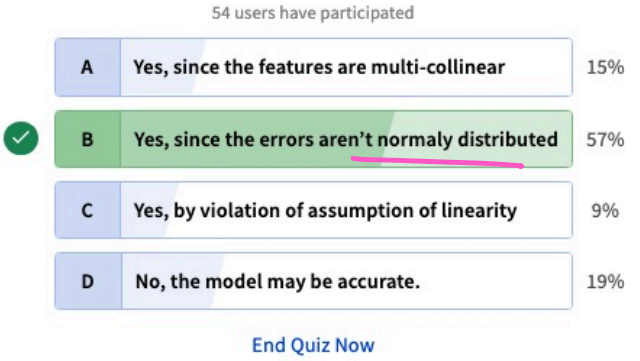


a few points with v. large the errors

"outliers"



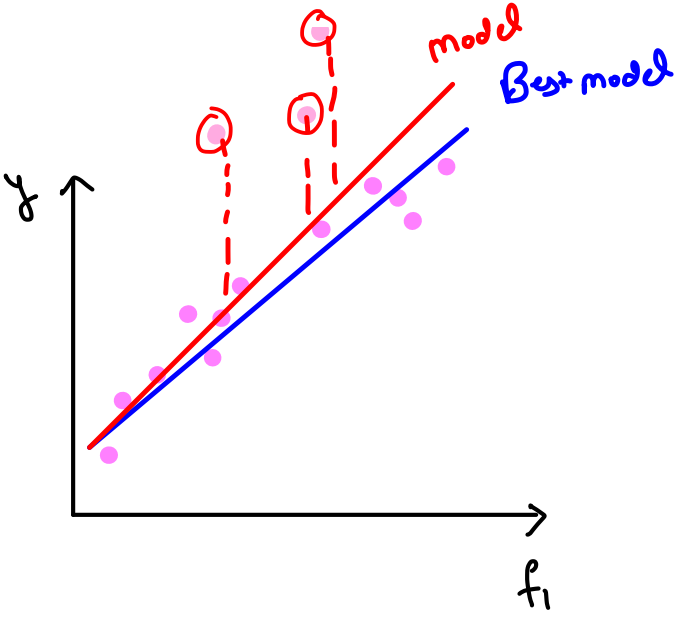
While building a risk prediction model for loan defaulters, it was observed that the erros were right skewed.
Does this imply anyway that the linear regression model is inaccurate?



Impact of outliers

I. identify outliers ?

II. Deal outliers !



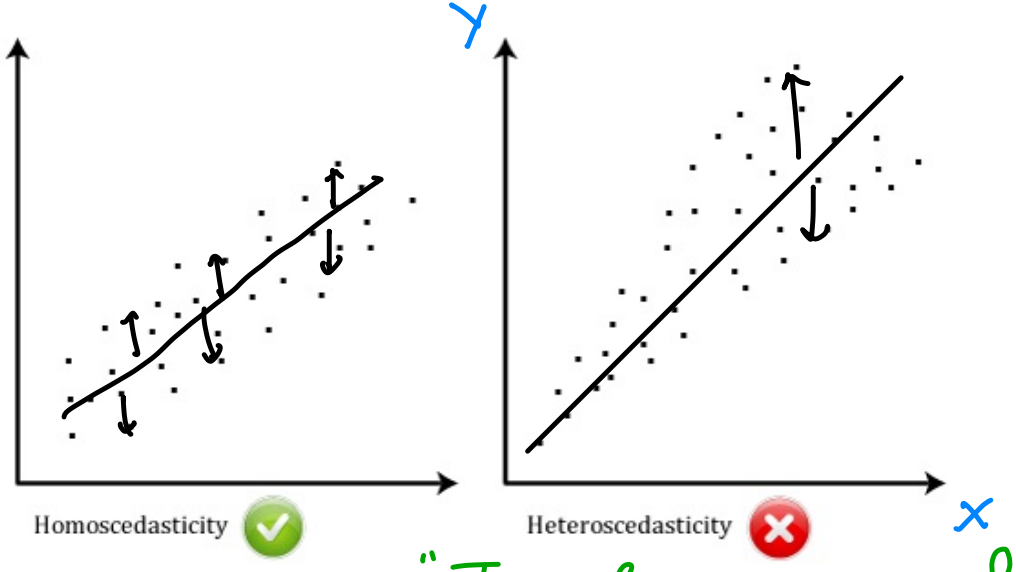
Residual Analysis

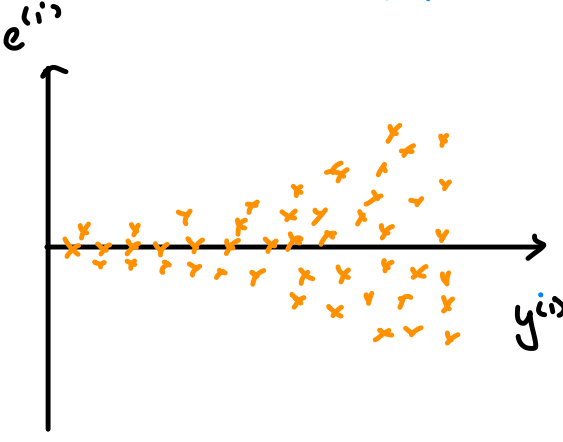
Error

No Heteroskadasticity

Homoskadasticity

$\epsilon^{(i)}$ vs. $y^{(i)}/\hat{y}^{(i)}$





"Transformation" → log transform

No. Autocorrelation

→ Time Series

Time	Date	Sales
-	-	
-	-	
-	-	
-	-	
-	-	
-	-	

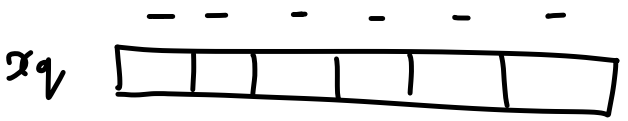
→ every data point should be independent

In linear regression, a high VIF value suggests:

38 users have participated

	A	Heteroskedasticity is present	11%
	B	A strong linear relationship between the independent and dependent variables.	11%
	C	The absence of outliers in the dataset.	0%
✓	D	Strong multicollinearity between predictor variables.	79%

End Quiz Now



"Scaling"

w_{Best}

$\hat{y} = w^T x_q + w_0$

model.predict(x_q)

$A^+ \rightarrow P(\text{Approved}) \quad 0.75$

$B^- \rightarrow P(\text{Approved}) \quad 0.61$