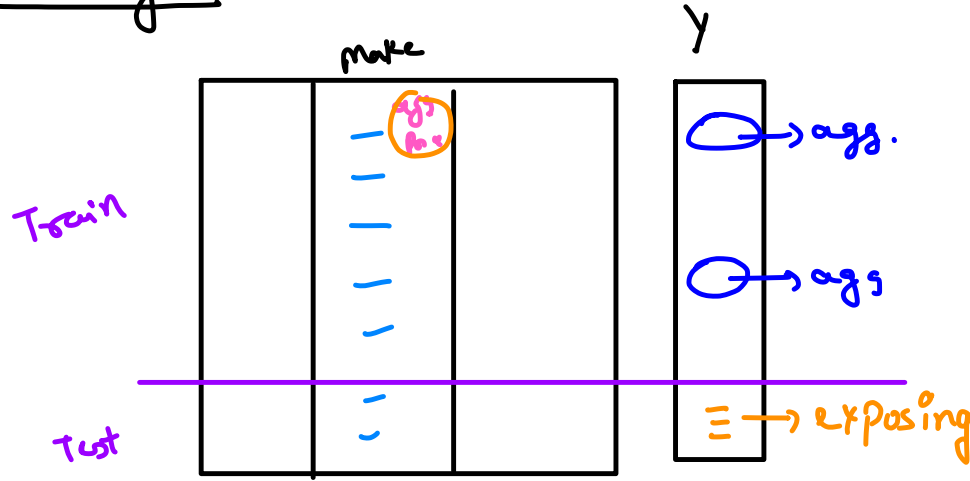


Linear Regression - 04

Data leakage



{ Maruti : 7.4
Honda : 8.1
TATA : 9.2
}

first Train/Test Split

Assumptions of Linear Regression

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

No Multicollinearity

Colinearity ?

f_1, f_2

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

f_1 & f_2 are colinear.

age, year

$$\text{age} = -1 \cdot \text{year} + 2025$$

Colinearity multiple feature \Rightarrow Multi-collinearity

f_1	f_2	f_3	f_4

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

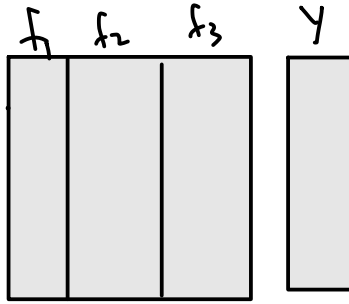
f_2 is multicollinearity

height (cm)		height (ft)
-------------	--	-------------

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

No Multicollinearity !!

Q → why MC is a Problem?



if m.c exists

$$f_2 = 1.5 x_1$$

optimized → $\omega^* = [\omega_1, \omega_2, \omega_3], \omega_0$

$$\omega^* = [1, 2, 3], \omega_0 = 5$$

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

$$\hat{y}_q = \omega_x^T x_q + \omega_0$$

$$\textcircled{1} - = 1 \cdot x_{q1} + 2x_{q2} + 3x_{q3} + 5$$

$$\omega = [1, 2, 3]$$

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

$$\textcircled{2} - = 4x_{q_1} + 3x_{q_2} + 5$$

$$w = [4, 0, 3]$$

$$x_q = [2, 3, 1]$$

$$\textcircled{1} \quad 1 \cdot 2 + 2 \cdot 3 + 3 \cdot 1 + 5$$

$$\begin{aligned} \hat{y} &= 2 + 6 + 3 + 5 \\ &= 16 \end{aligned}$$

→ No. feature importance

→ interpretability

$$\textcircled{2} \quad 4 \cdot 2 + 3 \cdot 1 + 5$$

$$\begin{aligned} \hat{y} &= 8 + 3 + 5 \\ \hat{y} &= 16 \end{aligned}$$

Messed up with weights → instability of weights.

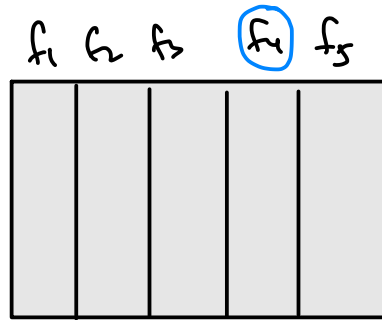
How does multicollinearity affect regression analysis ?

84 users have participated

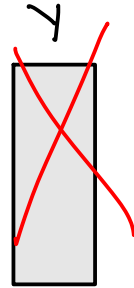


- | | | |
|---|---|-----|
| A | It reduces the interpretability of regression coefficients. | 77% |
| B | It increases the accuracy of the regression model. | 7% |
| C | It improves the goodness-of-fit of the regression model. | 4% |
| D | It has no impact on the regression analysis. | 12% |

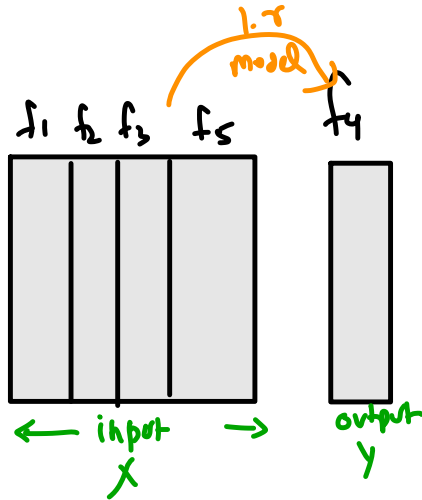
/// Variance Inflation factor [VIP]



← X →



VIF(f_4)



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

R^2 Score

if $R^2 = 0.18$ ∴ M.C Not exists

if $R^2 > 0.92$ ∴ M.C exists -

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

if $R^2 = 0$; $VIF \rightarrow 1$

if $R^2 \sim 1$; $VIF \rightarrow \infty$

$$VIF \in [1, \infty)$$

↓
No M.C

↘ M.C exists
v. large

$$VIF(f_1) = \text{—}$$

$$VIF(f_2) = \text{—}$$

⋮

$$VIF(f_d) = \text{—}$$

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



A **Linearity between the independent variables and sales.** 37%



B **Normal distribution of sales.** 3%

C **MultiCollinearity among features** 13%

D **All of them** 47%

Normality of Residuals

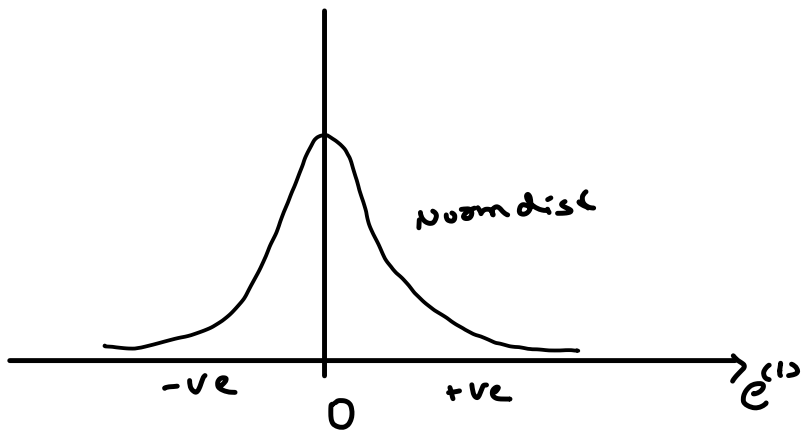
$$\rightarrow y - \hat{y}$$

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

$$\hat{y}$$

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

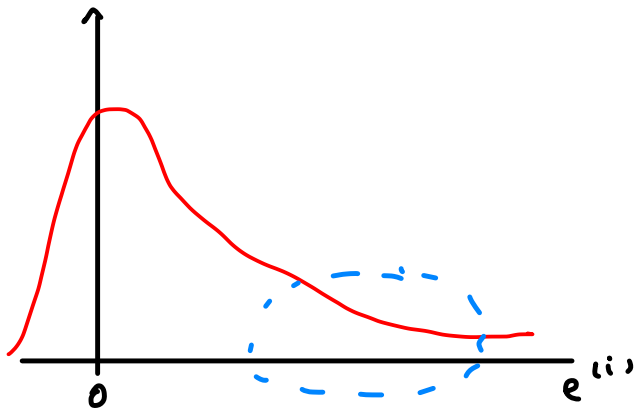
residuals



$$3.5L - 4L$$

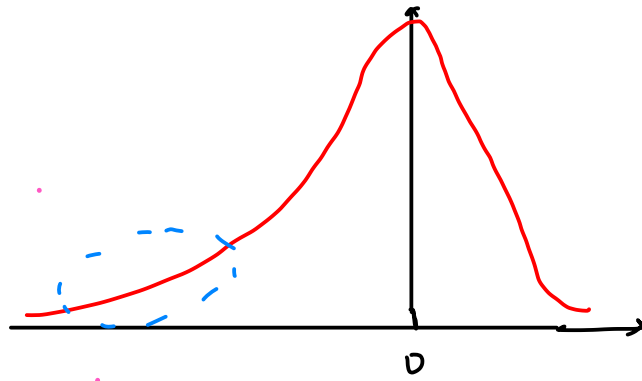
↓
-ve

→ over predicted



↓
 a few points
 with v. large τ
 errors

↓
 "Outliers"



While building a risk prediction model for loan defaulters, it was observed that the errors were right skewed.

Does this imply anyway that the linear regression model is inaccurate?

54 users have participated

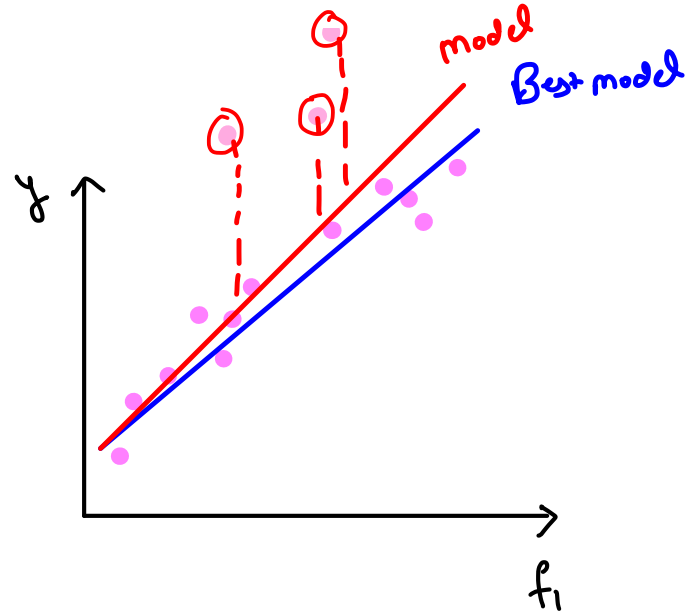
- | | | |
|-----|---|-----|
| A | Yes, since the features are multi-collinear | 15% |
| ✓ B | Yes, since the errors aren't normally distributed | 57% |
| C | Yes, by violation of assumption of linearity | 9% |
| D | No, the model may be accurate. | 19% |

[End Quiz Now](#)

Impact of outliers

- I. Identify outliers ?
- II. Deal outliers !

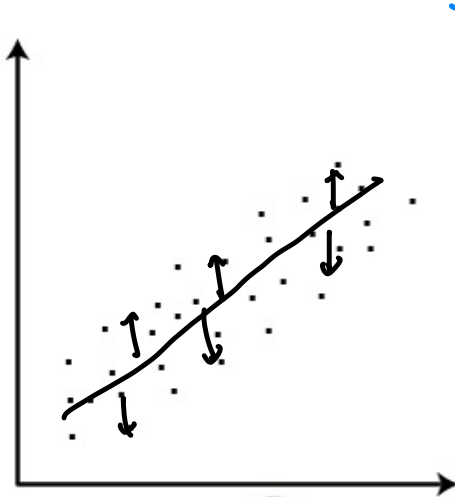
Residual Analysis
Error



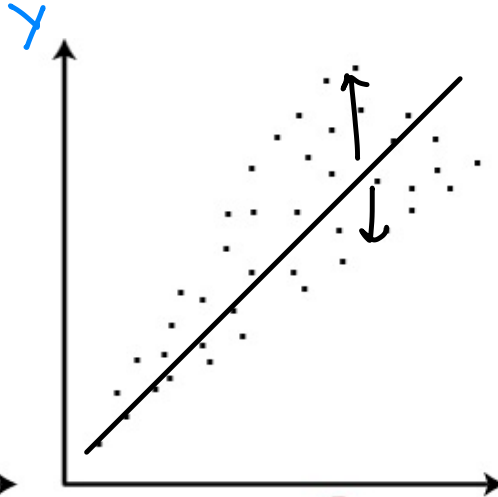


No Heteroskedasticity

Homoskedasticity



Homoscedasticity

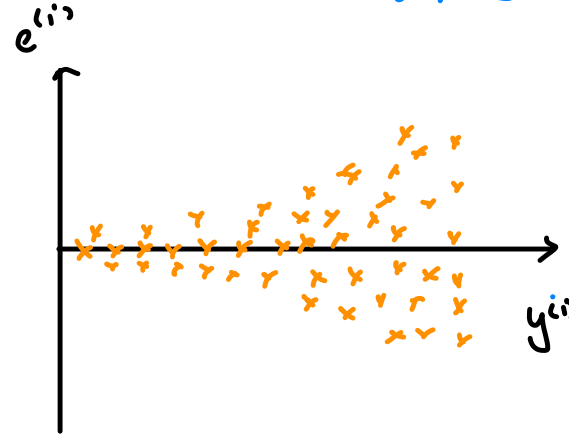


Heteroscedasticity



"Transformation" \rightarrow log transform

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





☒ No. Autocorrelation

Time Series

<u>Time</u>	<u>Date</u>	<u>Sales</u>
-	'	
-	-	
-	-	
-	-	
-	-	
-		<input type="checkbox"/>
-	-	
-	-	<input type="checkbox"/>

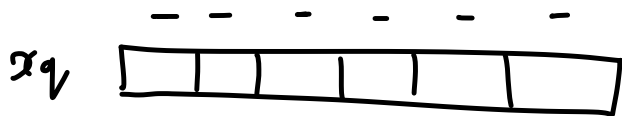
→ every data point should be independent

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"

$A^+ \rightarrow P(\text{Approved}) \ 0.75$
 $B^- \rightarrow P(\text{Approved}) \ 0.61$

$\hat{y} = w_{\text{best}}^T x_q + w_0$

model.predict(x_q)

Pincode			Loan / Reject
	A^+		1
	B^-		0
	B^-		1
	\vdots		1
	\vdots		0
	\vdots		0
	\vdots		0
	\vdots		0