

Assumptions of Linear Regression

• Set of conditions that if met, ensure that Linear Regression is a good choice and co-efficients (Parameter and intercepts) are reliable and Valid

- ① Assumption of Linearity
- ② No multi-collinearity
- ③ Normal Distribution of Residuals
- ④ Homoscedasticity
- ⑤ No auto-correlation

Assumption of Linearity

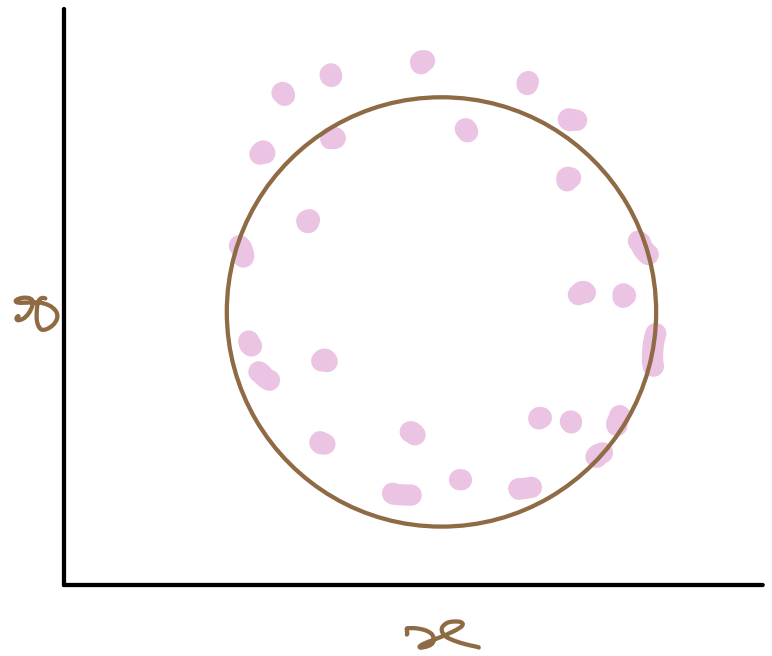
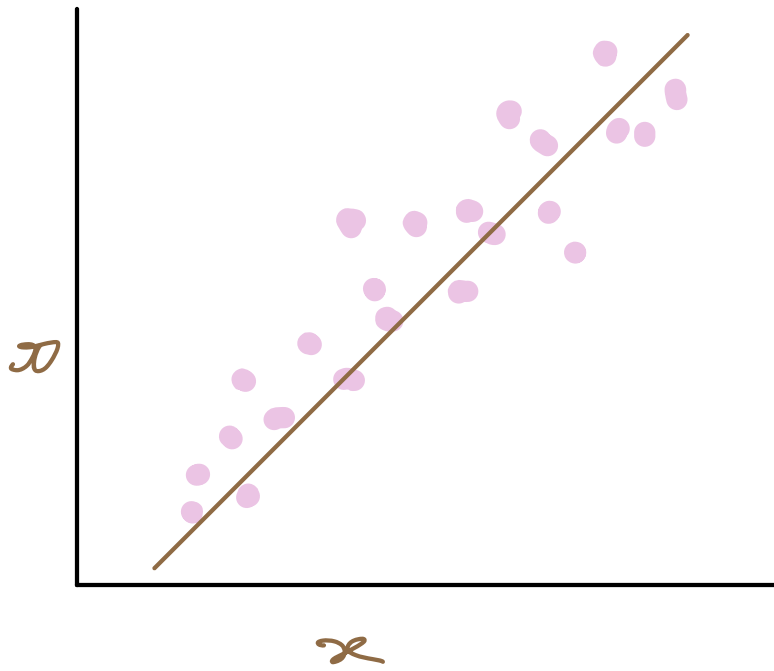
Definition: Target must have linear Relation with features

Questions How do we check linear Relationship?

Answers Pearson - correlation

$$x_1 \longrightarrow y$$

$$x_2 \longrightarrow y$$



Questions

If a feature is Not correlate we can drop or try F.E.

Questions

Linear Regression is Not Suitable for Non-Linear Relationship?

Answers

By default yes, How we can do Polynomial Regression and feature Engineering to convert Non Linear features \rightarrow Linear

To be Covered Later

No multi-collinearity

Definition: Features should Not be collinear with each other.

Collinear

$$y_1 = a + b y_2$$

Questions Why multi-collinearity is an issue?

Answers

LR
↓
Trained

$$\hat{y} = w_1 x_1 + w_2 x_2 + w_3 x_3 + w_0$$
$$\Rightarrow 1x_1 + 2x_2 + 3x_3 + 5$$

Let's assume linear Relationship

b/w 2 features

$$x_2 \Rightarrow 1.5 x_1$$

$$1x_1 + 2x_2 + 3x_3 + 5$$

↓

$$1x + 2(1.5x_1) + 3x_3 + 5$$

$$4x_1 + 3x_3 + 5$$

1
2
3

Same model
Contradictory

4
0
3

Questions How do we 'check' multi-collinearity?

Answers ① $f \rightarrow f'$ we can look at Heatmap and look for High value of correlations

$$f_1 = a(f_2) + b \quad (i)$$

$$f_1 = a(f_2) + b(f_3) + \dots + z \quad (ii)$$

$$\downarrow \qquad \qquad \downarrow$$
$$y = w_1 x_1 + w_2 x_2 + \dots \quad (ii)$$

② Build Linear Regression Model to predict $feature_{(i)}$ using all features as input

\downarrow
VIF

(Variance Inflation Factor)

Questions How do we 'Fix' multi-collinearity?

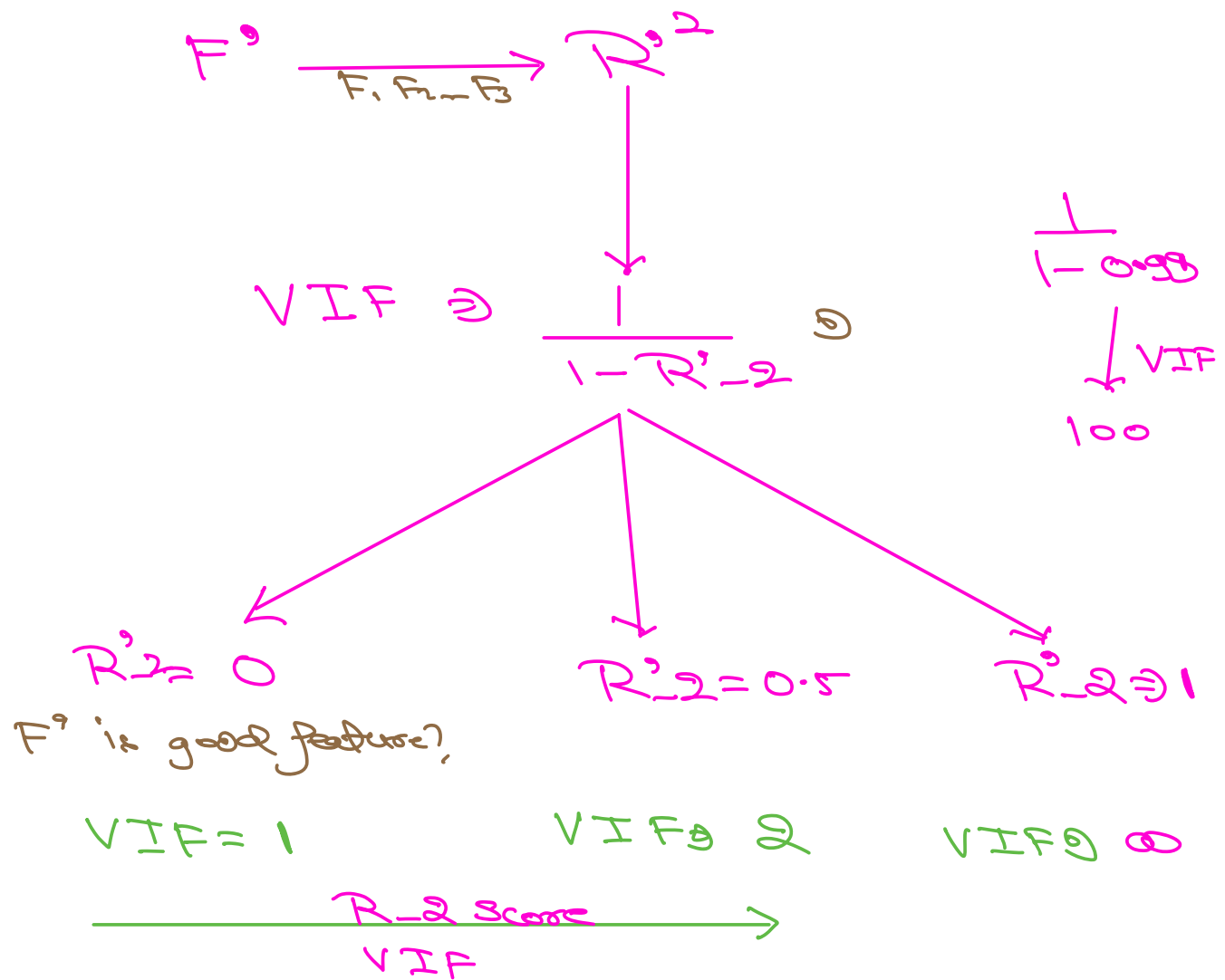
Answer: Drop features with high multi-collinearity

① Drop all features above certain correlation ± 0.90

② Calculate VIF and drop features one by one

					Target
F ₁	F ₂	F ₃	F ₄	-----	F ₉

In VIF \rightarrow one feature as T
 \rightarrow all other features as input



Higher VIF's are Bad and indicate multi-collinearity

- ① Calculate VIF for each of the features as target
- ② Check max VIF score, if $\max VIF > \text{threshold}(5)$ drop the max-feature
- ③ Recalculate and repeat until No $VIF > \text{threshold}$

4 features in total

$$\textcircled{1} \quad y_1 = a y_1 + b y_2 + c$$



$$y_1 = w_1 y_1 + w_2 y_2 + w_3 y_3 + c$$

↓
VIF
5.10

↓
 \mathbb{R}

↓
 \mathbb{R}

↓
5.0

$$y_2 = w_1 y_1 + w_2 y_2 + w_3 y_3 + c$$

↓
VIF
5.12

Q

Why not drop all feature with High VIF at once

Normality of Residuals

Definition: Residual must follow Normal Distribution

residuals/errors $\rightarrow y_i - \hat{y}_i$

$y \leftarrow \text{Dataset}$

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

Questions: How do we 'check' Normality of e ?

Answers:

- Plot distribution
 - Histogram
 - KDE plot
 - QQ plot

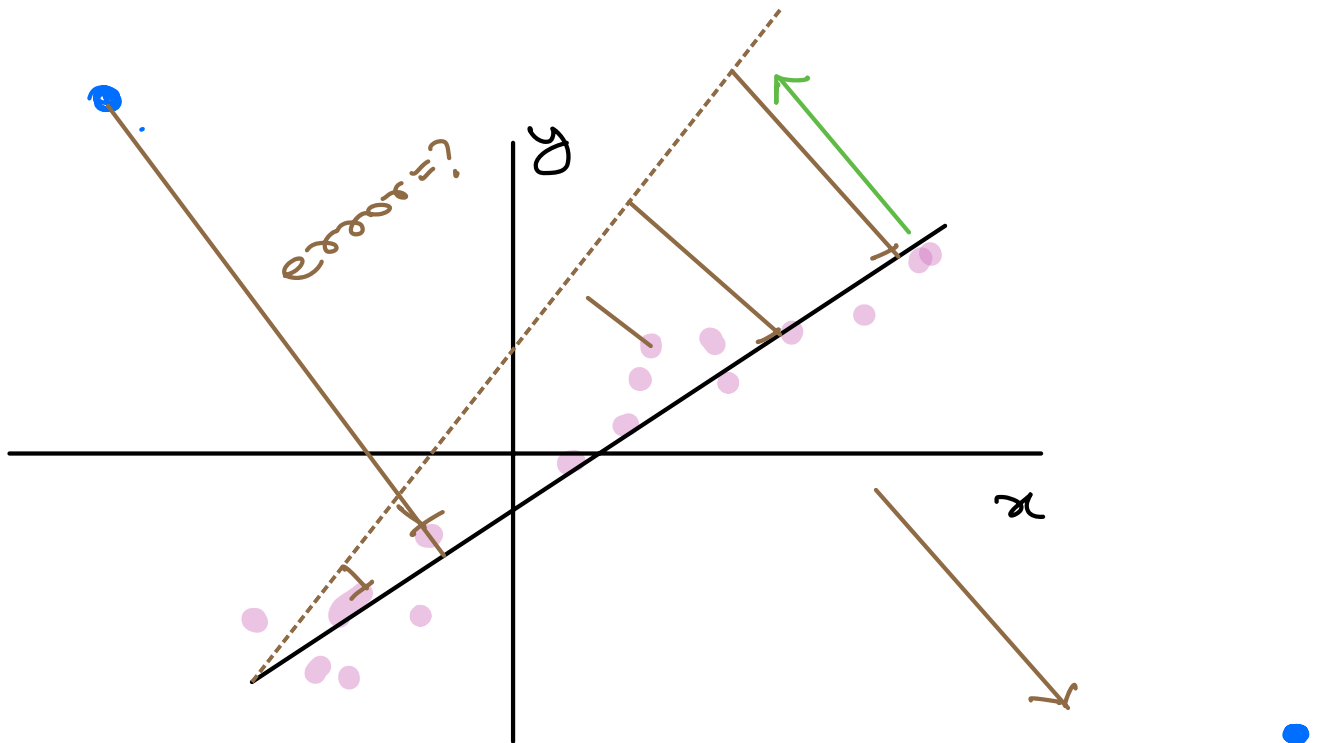
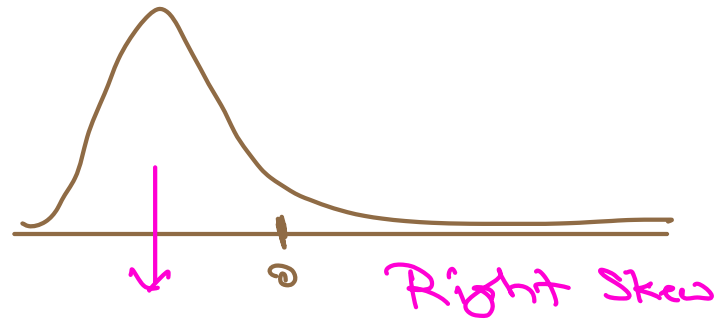
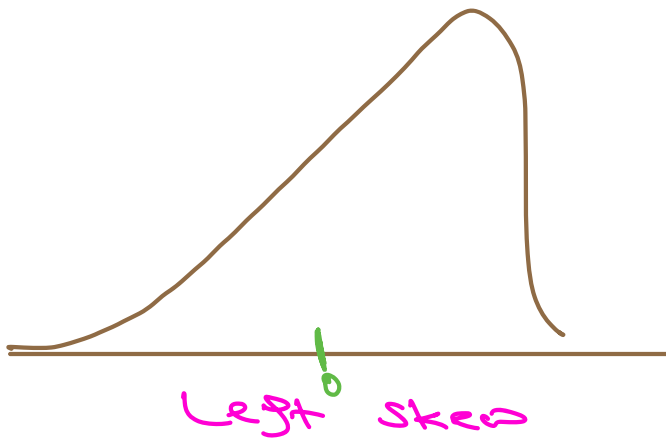
② Shapiro Wilkin test (errors)

H_0 : Sample comes from Normally distributed population

$$\text{Test_statistics} > 0.05 \quad \checkmark$$

0.05 reject H_0

Outliers in dataset



Questions How do we 'Fix'?

Answers Remove or impute Outlier

Homoscedasticity

No Heteroscedasticity

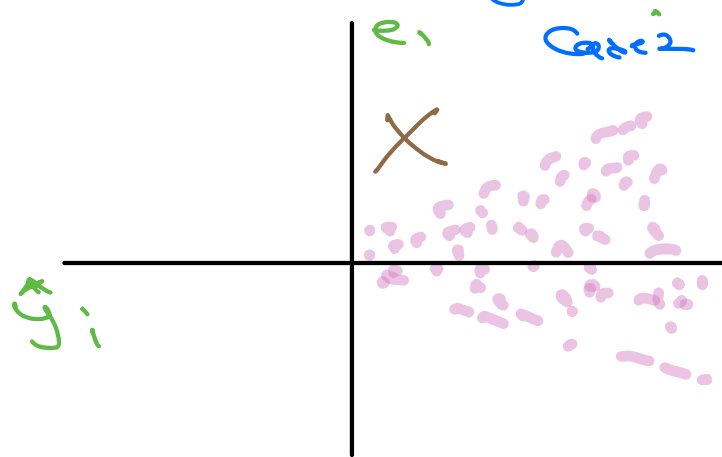
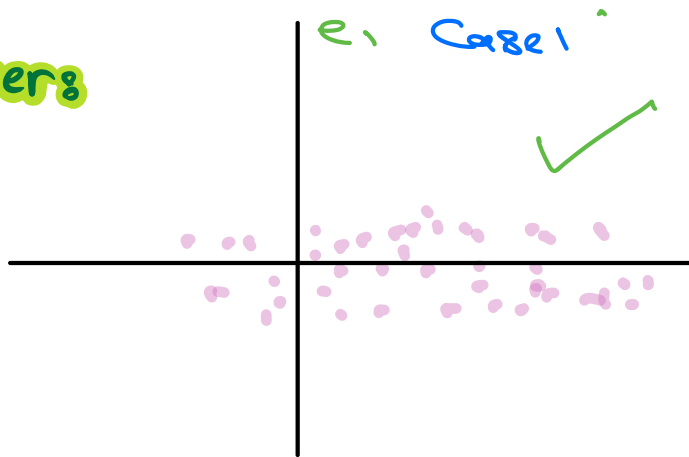
Definition:

$$\hat{y}_i \text{ vs } e_i$$

The variance of predictions against residual should not change

Questions: How do we check Homoscedasticity?

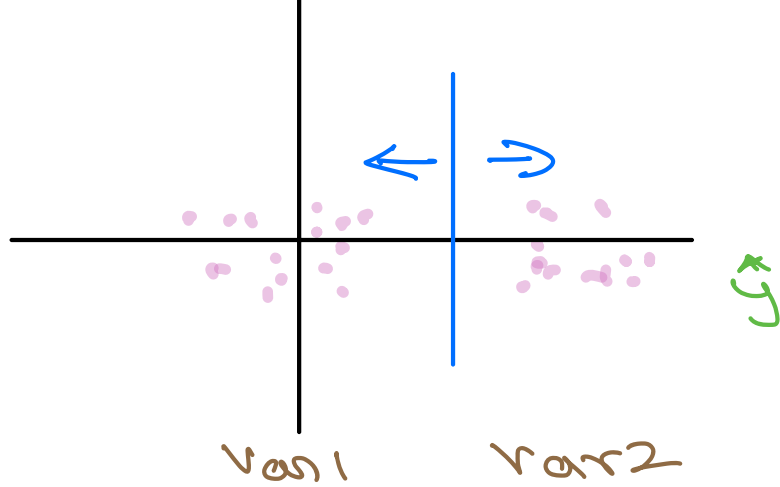
Answers:



Goldfeld Quandt test

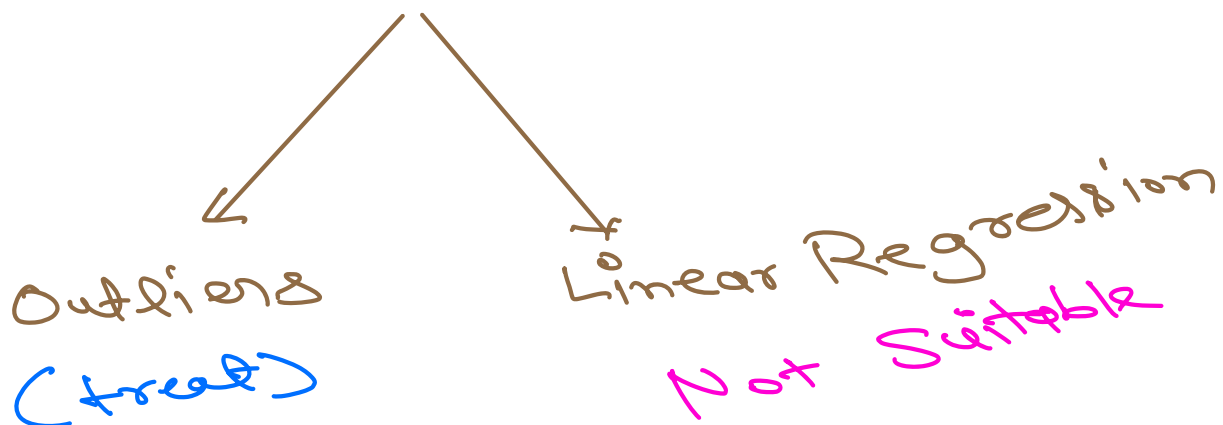
H_0 : Variance is Constant b/w two groups

H_a : Variance is Not Constant



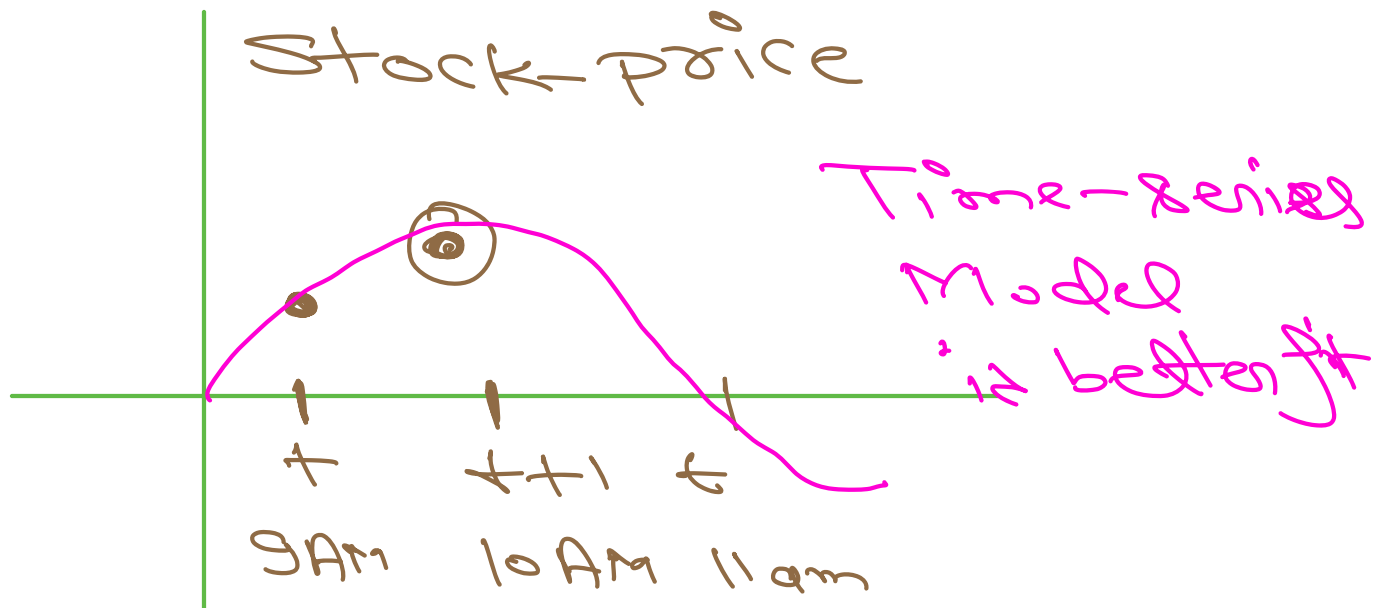
Questions How do we 'Fix'?

Answers If our errors are heteroscedastic



No AutoCorrelation

Where data is correlated with itself



Target variable is correlated with itself (Past Values)

Time-series Models

	0	y_0
	1	y_1
X	2	y_2
	3	y_3

14. Quiz 5

In linear regression, a high VIF value suggests:


 Launch

4 options

Active Duration (Most preferred: 30 seconds)

Appears for 45 Secs

A Heteroskedasticity is present

B A strong linear relationship between the independent and dependent variables. 

C The absence of outliers in the dataset.

D Strong multicollinearity between predictor variables.

$VIF \Rightarrow L(\text{feature})$

drop \rightarrow Target
 \downarrow
Dependent