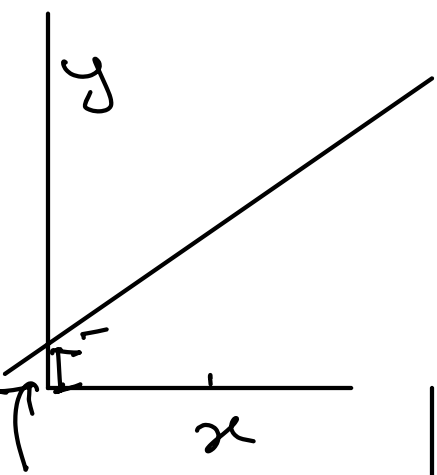
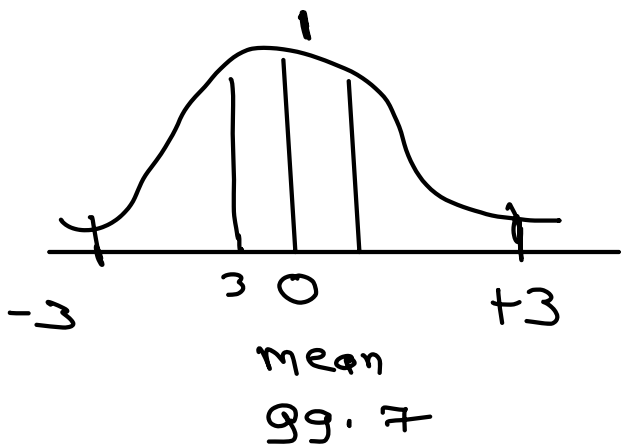
Linear Regression - 4

- Agenda
- Sklearn and statsmodel
- Assumptions of Linear Regression
  - Linear Relationship
  - No multi-collinearity
  - Normal Distribution of Residuals
  - Homoscedasticity
  - No auto-correlation





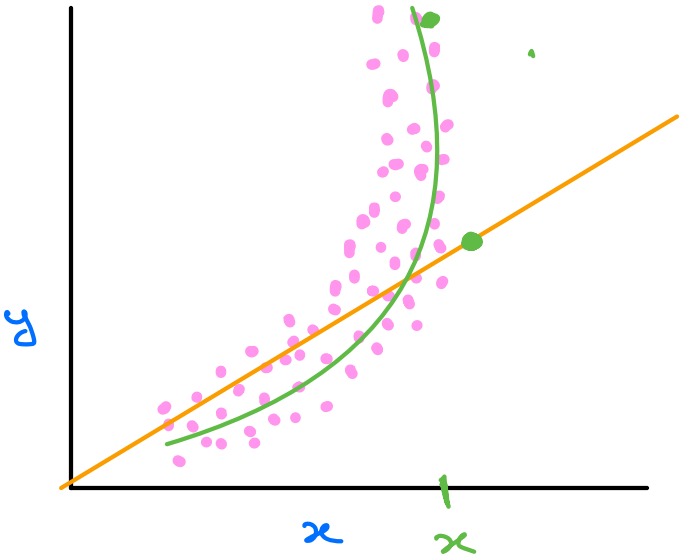
(1)

Training → 0.95
Test → 0.60
} overfitted on Training set

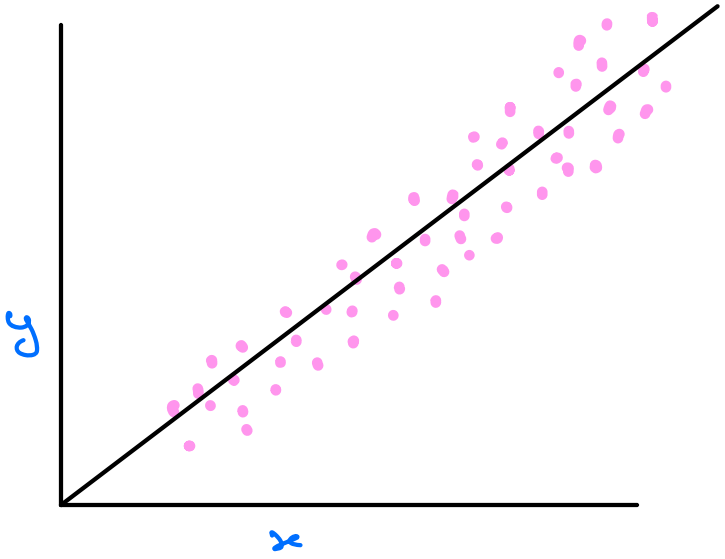
(2) Train 0.60
Test 0.60

0.95
0.85

Assumption of Linearity



Case -1



Case -2

1) Dependent Feature (y) should have linear Relationship with Independent Feature (x)

No multi-Collinearity

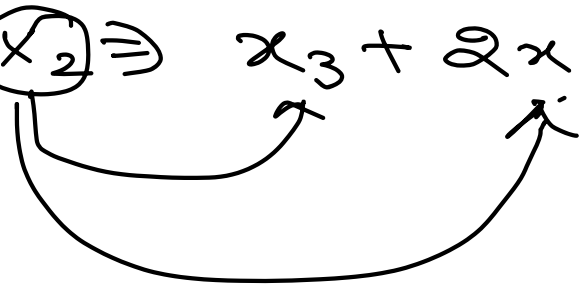
x_1	x_2	x_3	... x_d

y

=

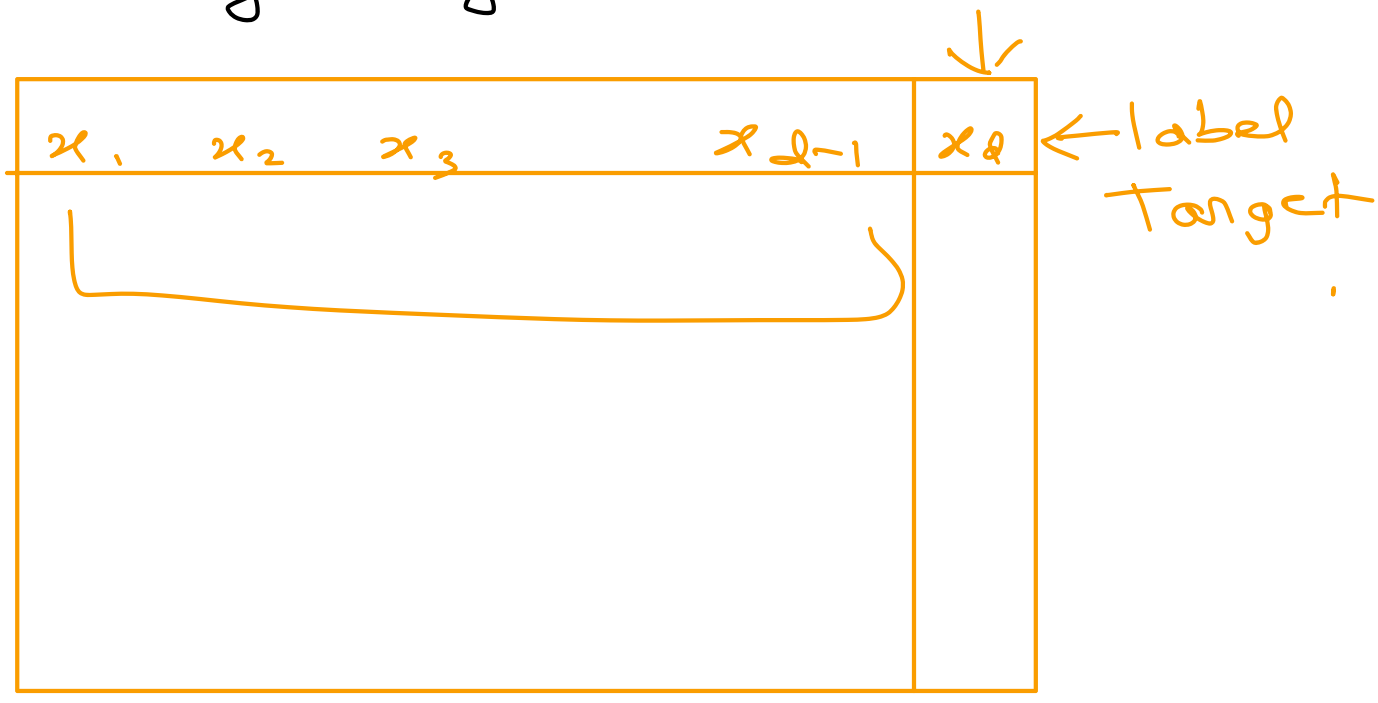
$x_1, x_2, x_3 \dots x_d$ should be independent of each other

case: $X_2 \Rightarrow x_3 + 2x_1 + 10$



year ⇒ feature

Age ⇒ feature

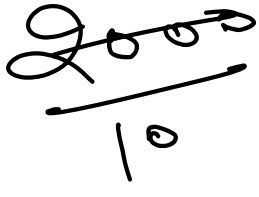


$r2\_score(X_d) \Rightarrow 0.01$

How Much x_d can be Explained with remaining features?

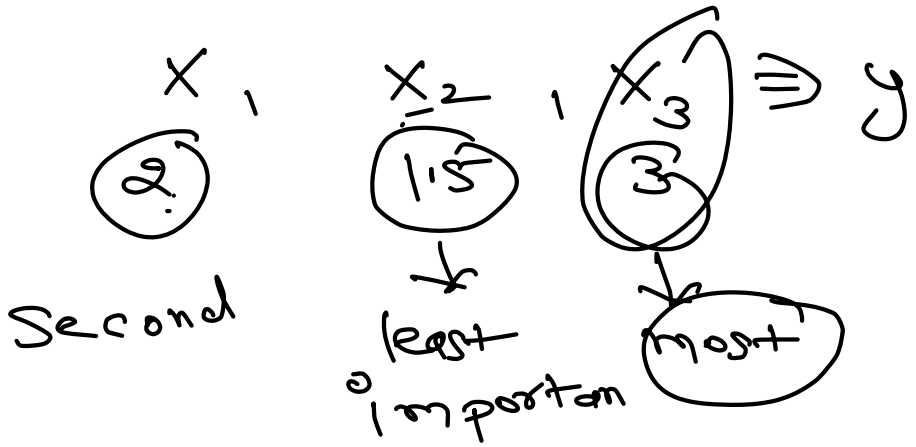
⇒ $X_1 \cdot X_2 = y$

$\Rightarrow$ 1 2



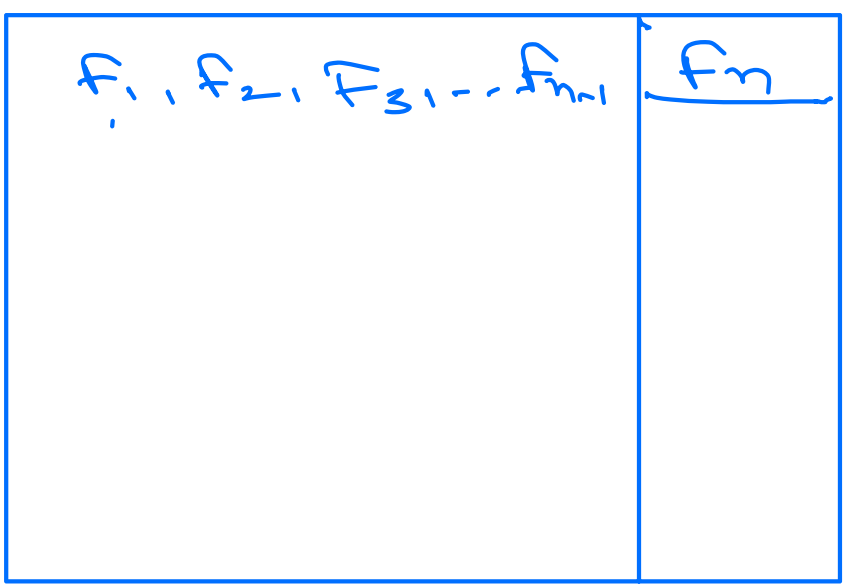
Age (10) ⇒ 20,000

2 (10+1)



$X_3 \Rightarrow \alpha X_1 + \beta X_2$

Variance_Inflation_Factor



6

Target

X ⇒ F1, F2, F3, F5, F6

Find R^2-score ⇒

$$VIF = \frac{1}{1-R_i^2}$$

$i = 0, 1, \ldots, n$

Variance Inflation factor

(a) $R_i^2 = 1$

$$VIF = \frac{1}{1-1} \Rightarrow \infty$$

(b) $R_i^2 = 1$

$$VIF = \frac{1}{1-0} \Rightarrow 1$$

⇒ VIF > 10 : Highly Multi-Coll (Drop it)

⇒ 5 ≤ VIF ≤ 10 : Collinear

⇒ VIF ≤ 5 (Good to keep)

How to drop

Step1: Build Regression model on All features

Step2 : Calculate VIF for each feature

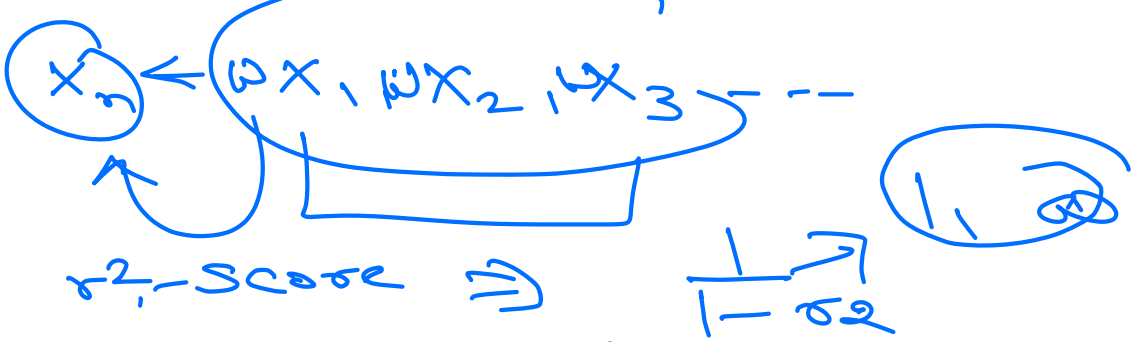
Step3 : Check and Drop feature with highest VIF

Step4 : Refit the model with remaing features

Step5 : Repeat until you have no feature with VIF>5

or

r2_score of target > threshold



r^2-score $\Rightarrow$ $\frac{1}{1-r^2}$

How x_n is dependent on multiple features

r^2_ Score = 0 $\quad w_1 = 0$

$w_2 = 0$

$w_{n-1} \to 0$

r^2 score = 1

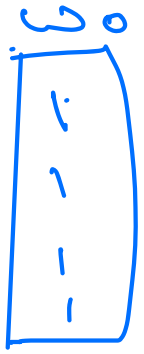
x_n can be explained / predicted by all other features

17 features + 1 w_0's and w_0

So

17 feature

add_Constant (X_Train)



17 + 1 = 18