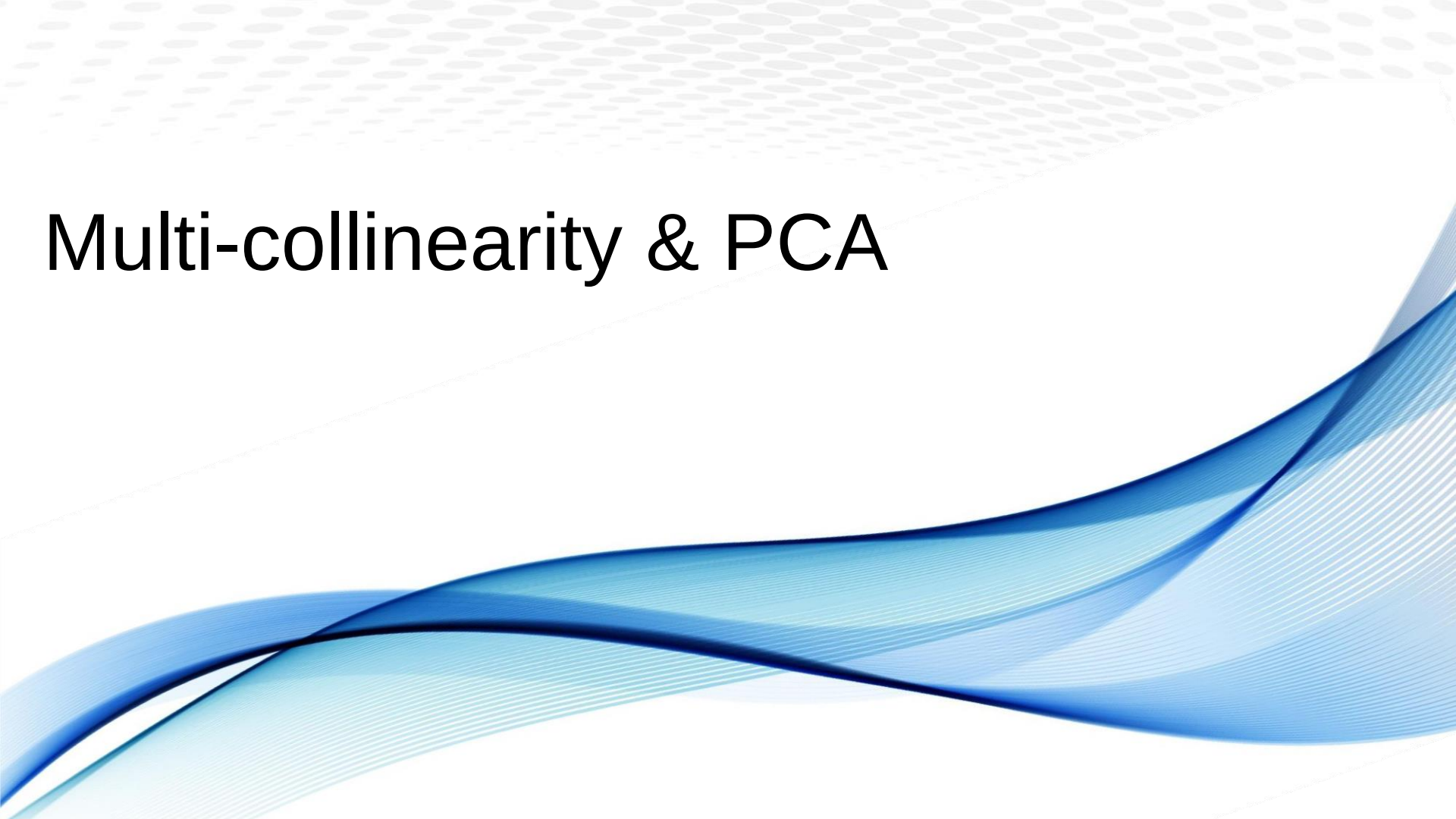
Multi-collinearity & PCA



Some important points to understand before proceeding

What are Independent & Dependent Variables?

A dependent variable denotes a number whose value is reliant on how the independent variable is modified/manipulated.

For example: in a credit card payment default problem, variable of interest will be default status (whether a person will default or not) or in loan applications data, variable of interest can be loan amount approved or approval status depending upon the business requirement.

What is a model?

In machine learning, a model is a set of steps based on mathematical/statistical concepts & assumptions to predict values of a dependent variable. Ex: Linear Regression Model

What are Predicted and Actual Values?

Predicted value is the result/output of the Model predictions. This may be either in 0s &1s or continuous or multi-label

The original value provided in the data set is Actual value. For example:

In a loan application, actual amount approved is $400,000 while your model predicted $340,500. Actual value is $400,000 and predicted value is $340,500.

Introduction

In Predictive Analytics, we try to establish a relationship between the Independent & Dependent Variables to get predictions

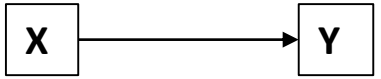


Fig 1

With only one independent variable and one dependent variable (fig 1)

With multiple independent Variables and a dependent variable (fig 2)

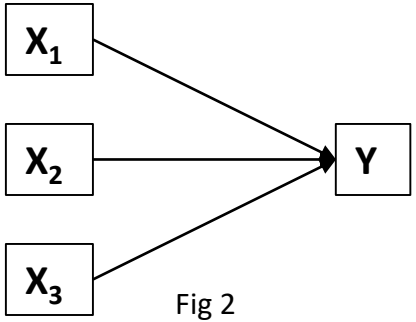


Fig 2

The difference in Actual Value and Predicted Value is called Error

- In Predictive Analytics, we try to establish a relationship between the Independent & Dependent Variables to get predictions

Fig 1

- With only one independent variable and one dependent variable (fig 1)

$\hat{Y} = a + bX$

- With multiple independent Variables and a dependent variable (fig 2)

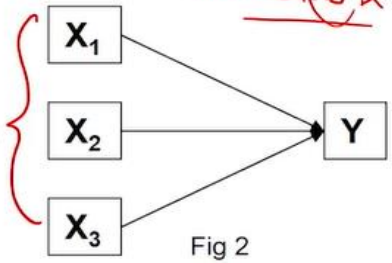


Fig 2

$\hat{Y} = a + b_1X_1 + b_2X_2 + b_3X_3$

- The difference in Actual Value and Predicted Value is called Error

$e = Y - \hat{Y}$ $\quad \Sigma e_i^2$

Correlation

Correlation gives the degree of relationship between two variables and is quantified by the correlation coefficient (Pearson's r) $Y = a + b_1 X_1 + b_2 X_2$

When the independent variables are highly correlated with each other, it is called Multi-collinearity

It becomes difficult for the model to determine the true effect of Independent variables on Dependent variable in case of high multi-collinearity

Example : Cost prediction models for 2-wheelers

Y = total cost X_1 = no. of engines

X_2 = no. of wheels

$$Y = 1,000 + 40 \times X_1 + 4 \times X_2$$

$$= 1,000 + 40 \times X_1 + 4 \times 2 \cdot X_1$$

$$= 1,000 + 40 X_1 + 8 X_1 = 1000 + 48 X_1 + 0 X_2$$

$X_2 = 2 X_1$

$$= 1,000 + 40 \times \frac{X_2}{2} + 4 X_2$$

$$= 1,000 + 20 X_2 + 4 X_2 = 1,000 + 0 X_1 + 24 X_2$$

$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$ ← coeff unstable

Solution - PCA

- PCA helps us address this problem of multicollinearity.
- The principal components are orthogonal to each other which means they are uncorrelated. Hence, multicollinearity is removed.

(1) Do not include correlated variables

2. $Y = \beta_0 + \beta_1 (PC_1) + \beta_2 PC_2$ ⇐ PC_1 & PC_2 are independent