DATA ANALYTICS

Unit 2:Introduction to Regression

Mamatha.H.R
Department of Computer Science and Engineering

DATA ANALYTICS

Unit 2:Introduction to Regression

Mamatha H R
Department of Computer Science and Engineering

What is Regression?

- Regression is a tool for finding existence of an association relationship between a dependent variable (Y) and one or more independent variables (X_1, X_2, …, X_n) in a study.
- The relationship can be linear or non-linear.
- A dependent variable (response variable) "measures an outcome of a study (also called outcome variable)".
- An independent variable (**explanatory variable**) "explains changes in a response variable".
- Regression often set values of explanatory variable to see how it affects response variable (predict response variable)

Regression

Regression is a **supervised learning algorithm** under Machine Learning terminology

An important tool in **Predictive Analytics**

Regression model establishes existence of association between two variables, but not causation.

BCCI Bans Girlfriends and Wives

Girlfriends and wives create such a “distraction” that Indian batsmen can’t make runs, bowlers fail to take wickets and fielders drop simple catches.

Regression is not designed to capture causal relationship

Dependent and Independent are just terms used!

Married Men Earn More Money!

Is marriage leading to more money or More Money leading to Marriage?

Regression helps to validate hypothesis

Interesting Hypotheses

- Good looking couples are more likely to have girl child(ren)!
- Married people are more happier than singles!!!
- Vegetarians miss fewer flights.
- Black cars have more chance of involving in an accident than white cars in moon light.
- Women use camera phone more than men.
- Left handed men earn more money!
- Smokers are better sales people.
- Those who whistle at workplace are more efficient.

Regression - Definition

A statistical technique that attempts to determine the existence of a possible relationship between one **dependent variable** (usually denoted by Y) and a collection of Independent variables.

Regression is used for generating new hypothesis and for validating a hypothesis

Dependent and Independent Variables

- Terms dependent and independent does not necessarily imply a causal relationship between two variables.

- Regression is not designed to capture causality.

- Purpose of regression is to predict the value of dependent variable given the value(s) independent variable(s)

DATA ANALYTICS

Regression Nomenclature

Dependent Variable	Independent Variable
Explained Variable	Explanatory variable
Regressand	Regressor
Predictand	Predictor
Endogenous Variable	Exogenous Variable
Controlled Variable	Control Variable
Target Variable	Stimulus Variable
Response Variable	
Feature	Outcome Variable

Regression Vs Correlation

- Regression is the study of, "existence of a relationship", between two variable. The main objective is to estimate the change in mean value of independent variable.

- Correlation is the study of, "strength of relationship", between two variables.

DATA ANALYTICS

Where is it used?

- ✓ Every functional area of management uses regression.
- ✓ Finance: CAPM, Non-performing assets, probability of default, Chance of bankruptcy, credit risk.
- ✓ Marketing: Sales, market share, customer satisfaction, customer churn, customer retention, customer life time value.
- ✓ Operations: Inventory, productivity, efficiency.
- ✓ HR – Job satisfaction, attrition.

Importance of Regression

- In 1980, Supreme court of USA recognized regression as a valid method of identifying discrimination.

- American Food and Drug Administration (FDA) uses regression as an approved tool for validating food and drug products.

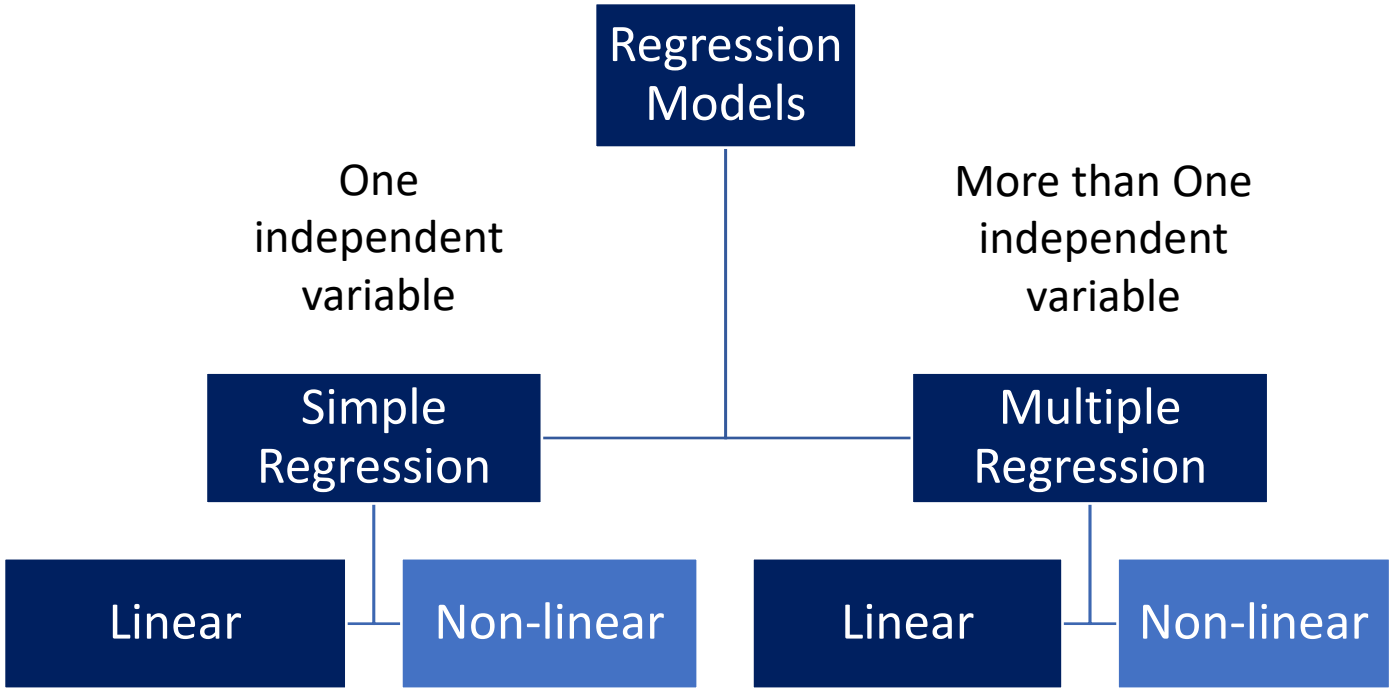
Why we need Regression?

- Companies would like to know about factors that has significant impact on their Key Performance Indicators (KPI).

- Regression helps to create new hypothesis that may assist the companies to improve their performance.

DATA ANALYTICS

Types of Regression



Types of Regression

- Simple linear regression – refers to a regression model between two variables.

$$Y = \beta_0 + \beta_1 X_1 + \varepsilon$$

- Multiple linear regression – refers to a regression model on more than one independent variables.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \varepsilon$$

- Nonlinear regression.

$$Y = \beta_0 + \frac{1}{\beta_1 + \beta_2 X_1} + X_2^{\beta_3} + \varepsilon$$

Exercise

- Bring out the difference between Regression and Correlation with some examples

References

Text Book:

"Business Analytics, The Science of Data-Driven Decision Making", U. Dinesh Kumar, Wiley 2017

THANK YOU

Dr.Mamatha H R

Professor,Department of Computer Science

mamathahr@pes.edu

+91 80 2672 1983 Extn 834