30 Hours Online Certificate Course On *"Research & Data Analysis"*

- Organised By: **Accounting and Finance Lab, Department of Commerce, Ramanujan College, University of Delhi**
- Resource Person: **Dr. Arnav Kumar**
- Designation: Assistant Professor, Department of Management Studies, Ramanujan College
- Email: arnavkumardse@gmail.com
- Day & Date: **June 17, 18, 19, 20 & 21, 2020.**

Non Parametric Tests-Methods

- A parametric statistical test is one that makes assumptions about the parameters of the population from which one's sample is drawn, while a non-parametric test is one that makes no such assumptions.
- Thus, Non-Parametric tests are a small family of tests that can be used to test hypotheses but don’t make many assumptions.
- Non-Parametric tests are also called or ‘Assumption-Free Tests’ or “Distribution Free Tests” because they don't assume that your data follow a specific distribution especially the assumption about normally distributed data.

Non Parametric Tests-Methods

- When to use Non-Parametric Tests?
- ✓ If we use a parametric test and a non-parametric test on the same data, and those data meet the appropriate assumptions, then the parametric test will have greater power to detect the effect than the non-parametric test.
- ✓ But Non-Parametric tests have less power only if the sampling distribution is normal.
- ✓ Hence, when Assumptions of other Parametric tests are not met, better use Non Parametric test.
- ✓ Also, they are used when no Robust methods are available in case of Violation of Assumptions.

Parametric Vs Non Parametric Tests

	Parametric	Non-parametric
Assumed distribution	Normal	Any
Assumed variance	Homogeneous	Any
Typical data	Ratio or Interval	Ordinal or Nominal
Data set relationships	Independent	Any
Usual central measure	Mean	Median
Benefits	Can draw more conclusions	Simplicity; Less affected by outliers
Tests		
Choosing	Choosing parametric test	Choosing a non-parametric test
Correlation test	Pearson	Spearman
Independent measures, 2 groups	Independent-measures t-test	Mann-Whitney test
Independent measures, >2 groups	One-way, independent-measures ANOVA	Kruskal-Wallis test
Repeated measures, 2 conditions	Matched-pair t-test	Wilcoxon test
Repeated measures, >2 conditions	One-way, repeated measures ANOVA	Friedman's test

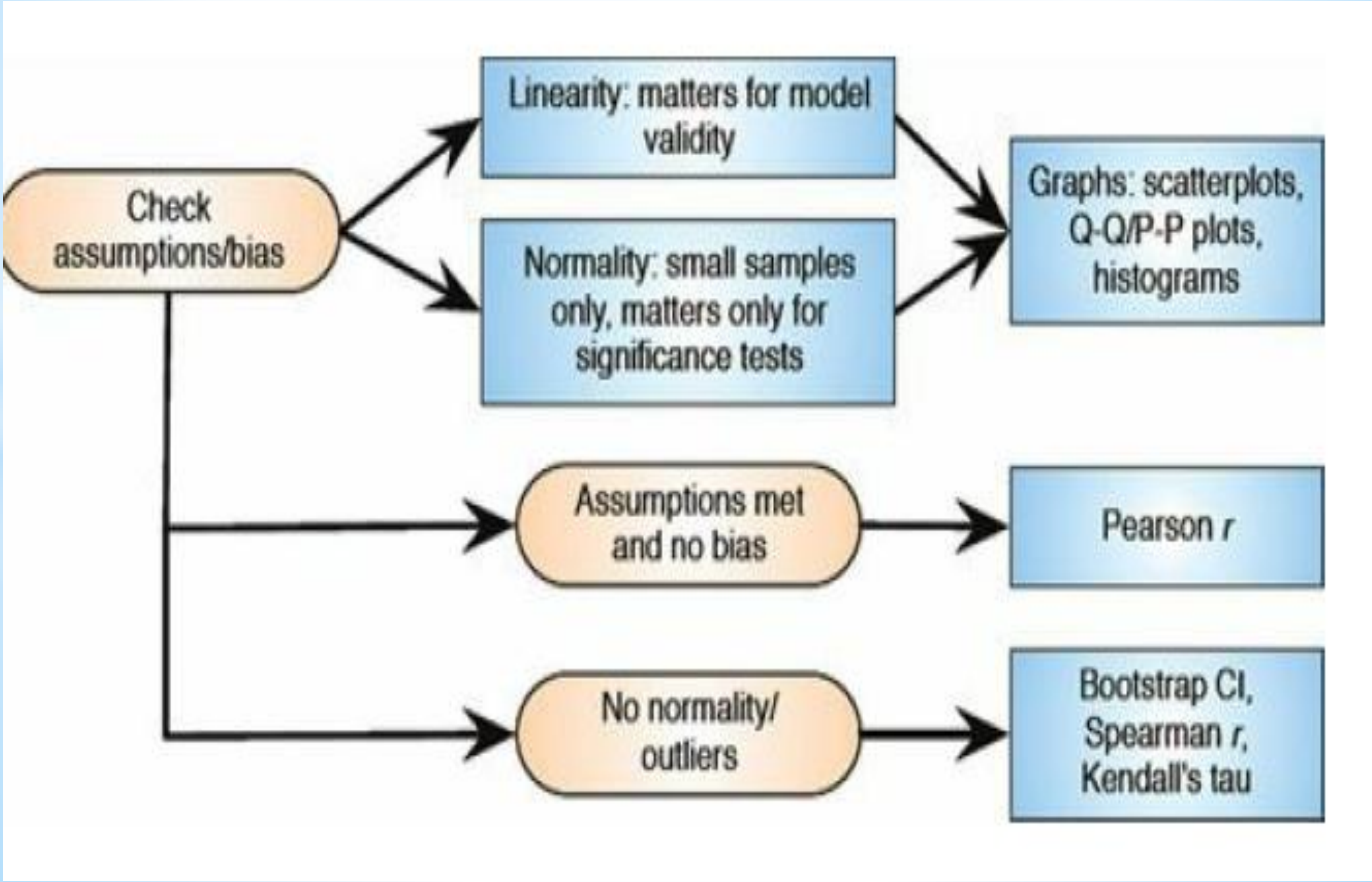
Correlation Analysis

Correlation Analysis

- Strength of Association between Variables
- Nature of Relationship: Positive or Negative
- Correlation Coefficient: -1 to +1.
- Perfectly Positive (+1) or Perfectly Negative (-1) or No Relationship (0).
- What if Correlation Coefficient is 0?
- What is the unit of Correlation Coefficient?
- Magnitude of Relationship
- ✓ High: More than 0.7 or Less than -0.7
- ✓ Medium: 0.4 to 0.7 or -0.7 to -0.4
- ✓ Low: Less than 0.4 or More than -0.4

Correlation Analysis

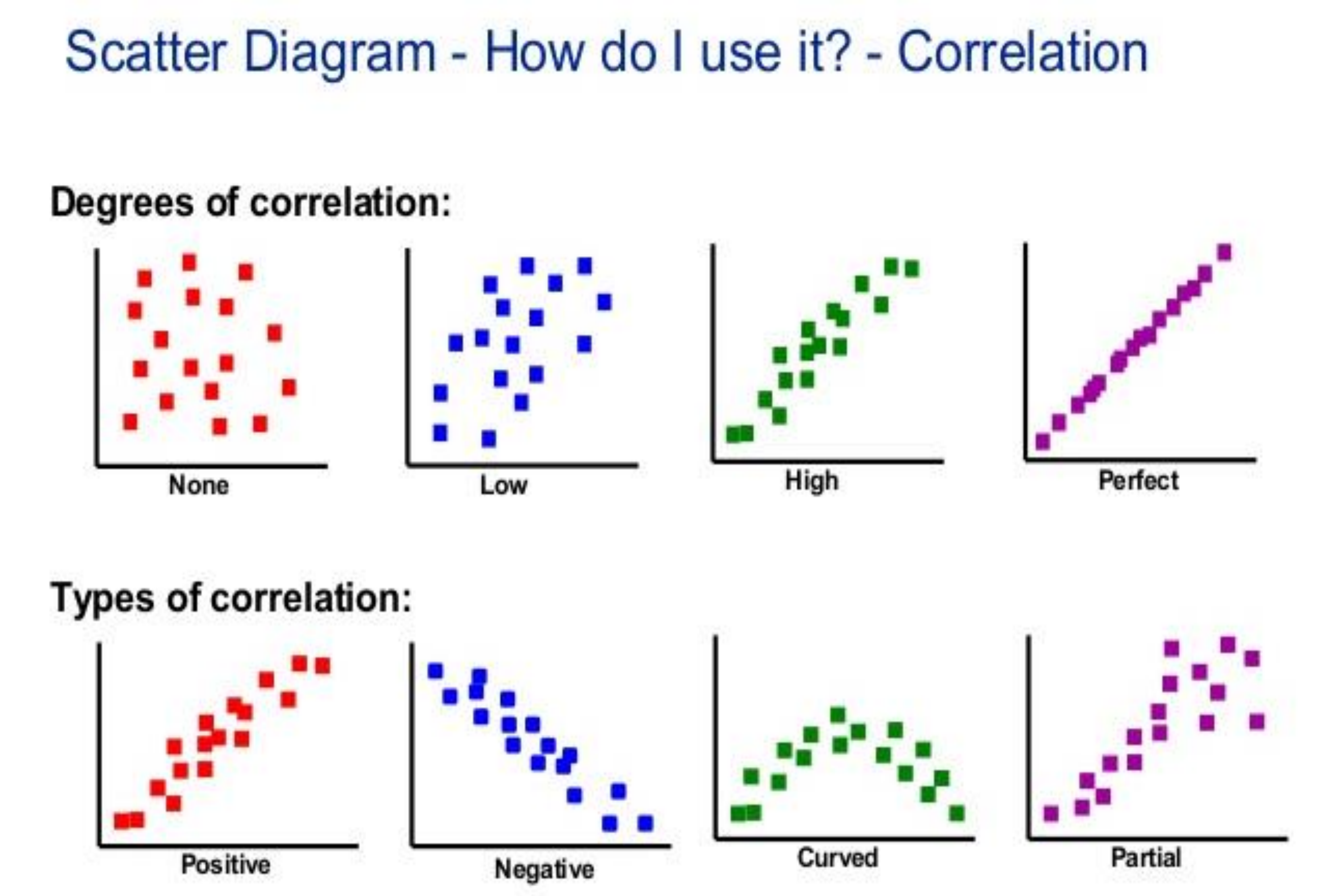
- **The General Process for Correlation Analysis**



Graphing Relationships: Scatterplot

- A Scatterplot is a graph that plots each case/respondent's score on one variable against their score on another variable.
- It tells us whether there seems to be a Relationship between the Variables, what kind of relationship it is and whether any cases are markedly different from the others.
- Simple Scatter: Plots values of one Continuous Scale Variable against another.
- Grouped Scatter: This is like a simple scatterplot, except that we can display points belonging to different groups in different colours (or symbols).

Graphing Relationships: Scatterplot



Graphing Relationships: Scatterplot

- **Simple Scatter Plot** in IBM SPSS Statistics:
- In Chart Builder, click Gallery & select Scatter/Dot in the Choose From list.
- Drag the Simple Scatter icon onto the canvas.
- Drag a Scale Variable (*GPA*) to X-axis.
- Drag another Scale Variable *(Percent)* to Y-axis.
- There is no need to specify a statistic, because Scatter Plots typically display raw values.
- Fit Regression Line: Double Left Click to open Chart Editor > Elements > Fit Line at Total.
- Also fit a 95% Confidence Interval around the Regression Line (Choose "Interval").

Types of Correlation

- **Simple/Bivariate Correlation:** Simple Correlation coefficient between two variables without considering affect of any other variable.
- **Partial Correlations:** Partial Correlations procedure computes partial correlation coefficients that describe the linear relationship between two variables while controlling for the effects of one or more additional variables.
- **Pearson's Correlation Coefficient**, is a parametric statistic & requires interval data for both variables. To test its significance, we also assume normality. We use Bootstrap for calculating Pearson Correlation Coefficient, if Normality is unsure.

Types of Correlation

- **Spearman's Rank Correlation Coefficient**, r_s, is a non-parametric statistic and requires only ordinal data for both variables.
- **Kendall's correlation coefficient, τ**, is like Spearman's r_s but is better for small samples.
- Null Hypothesis (Ho) for all Correlation Tests: Correlation Coefficient = 0, i.e., No Relationship.
- We can also do One Tail or Two Tail Tests in SPSS.
- **SPSS: Correlation Analysis**

Analyze > Correlate > Bivariate (*Final-Quiz 1*).

Analyze > Correlate > Partial (*Final-Quiz 1; Control for Quiz 2, Quiz 3, Quiz 4 & Quiz 5*).

Regression Analysis

Regression

- Regression is a statistical measure that attempts to determine the strength of relationship between a Dependent Variable (Y) and one or more Independent Variables (X).
- Regression is the backbone, corner stone & central theme of modern Statistics.
- It is the most widely used Statistical technique, being used my Academicians & Researchers, Businesses, Government Organisations, Research Institutions.
- It is heavily used in all disciplines of Business, Management, Sciences, & Social Sciences.

Importance & Uses of Regression

- To establish Relationship between Dependent and Independent Variables in the Population from Sample Data.
- To gauge the Cause and Effect Relationship between Dependent & Independent Variables.
- To obtain the value of Population Parameters.
- To know the Explained & Unexplained Variations.
- To Minimise Unexplained Variations (Error term).
- To estimate Predicted Value of Dependent Variable from Model using Independent Variables.
- To allow Forecasting – Within & Outside Sample.

Correlation, Regression & Causation

- **Correlation Coefficients** give no indication of direction of causality. There are two problems:
- **The Third Variable Problem or Tertium Quid**: In any Correlation, causality between two variables cannot be assumed because there may be other measured or unmeasured variables affecting the results.
- **Direction of Causality**: Correlation coefficients say nothing about which variable causes the other to change.
- Whereas, a **Regression Model** clearly implies Causation from the Independent Variable(s) to the Dependent Variable in the Model.

Dependent & Independent Variable(s)

S.No.	Dependent Variable	Independent Variable
1.	Explained	Explanatory
2.	Predictand	Predictor
3.	Regressand	Regressor
4.	Effect	Cause
5.	Outcome	Co-Variate
6.	Response	Stimulus
7.	Controlled Variable	Control Variable
8.	Endogenous	Exogenous

Regression Equation

- A Simple Regression Model is: $\mathbf{Y_i = b_o + b_1X_i + e_i}$
- ✓ Y_i = The ith Observation/Case of Regressand Y;
- ✓ X_i = The ith Observation/Case of Regressor X;
- ✓ b_o = Intercept Parameter which tells the value of the Outcome when the Predictor is zero;
- ✓ b_i = Slope Parameter quantifies the relationship between Predictor & Outcome. Denotes Sign (+Ve or –Ve), Magnitude, & Significance of Relationship.
- ✓ e_i = Error or Residual Term for the ith Case. It accounts for all other Variables (which are not used in the Model) impacting the Regressand (Y).
- ✓ Error(e_i) = $\mathbf{Y_i - (b_o + b_1X_i)}$ = **Actual Y - Predicted Y.**

Regression Equation

- Parameters (b_0, b_1) are also called Regression Coefficients.
- Linear equation here simply means a Straight Line which describes the relationship between Dependent & Independent Variables.
- We can use a linear model (i.e., a straight line) to summarize the relationship between two variables:
- ✓ Gradient or Slope (b_1) tells us what the model looks like (its shape); and
- ✓ Intercept (b_0) tells us where the model is (its location in geometric space).

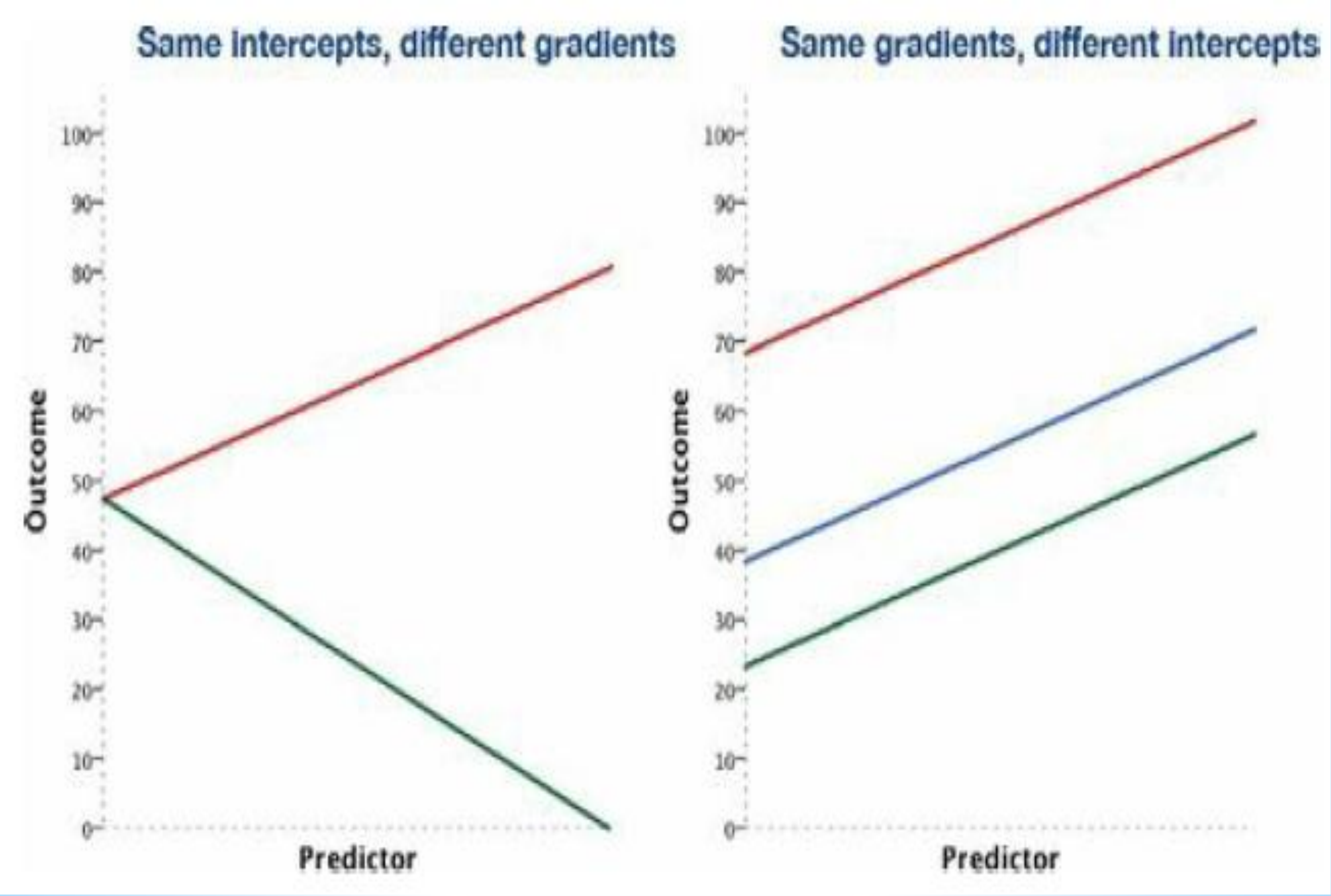
Linear Regression Model

- What is Linear? - Any Variable or Parameter is "Linear" if:
- If it appears with a Power or Index of 1 only.
- ✓ Anything with Power > 1 (X^2 or X^3) is Non Linear;
- ✓ Anything with Power < 1 ($X^{1/2}$ or $X^{1/3}$) is Non Linear;
- If it is not Multiplied or Divided by any other Variable. However, following is still Linear:
- ✓ If added or subtracted by any other Variable;
- ✓ If Multiplied or Divided by a Constant like 4X.

Linear Regression Model

- Linear Regression Model means that the Population Parameters (b_o, b_1) that are to be estimated from Sample data must be Linear.
- So, Parameters must have Power of only 1 & not be divided or multiplied by any variable.
- Even if the Independent Variables (Xs) are Non Linear, Regression will be Linear as long as its Parameters (b_o, b_1) are Linear.
- Linear Regression estimates the Coefficients of a Linear equation, involving one or more Independent Variables, that best predict or explain the value of Dependent Variable.

Linear Regression Model



Statistical Modelling

- **Statistical Model:** Outcome = Model + Error. Simplest Statistical Model is Mean.
- **Error =** Actual/Observed/Outcome value of a Variable – Predicted Value by Model.
- **Total Error** = Sum of Errors = ∑(Actual – Predicted).
- **Sum of Squared Errors (SSE)** = ∑(Actual – Predicted)^2
- **Mean Squared Error (MSE)** = Sum of Squared Errors (SSE)/Degrees of Freedom = ∑(Actual – Predicted)^2/(N – K).
- **Lower the MSE, better the fit of the Model.**

Estimating Population Parameters from Sample – The Method of Least Squares

- Regression Models are defined by Parameters, and these parameters need to be estimated from the data that we collect.
- Population Parameters are Unknown to begin with. So, we need some Procedure/Method to estimate them from Sample Data.
- Lower the MSE, better the fit of the Model.
- So, we use Method of Least Squares or Ordinary Least Squares (OLS) which estimates those values of Parameters (b_o, b_1) which minimises MSE of model & ensures Best Fit.

Simple and Multiple Regression

- Simple Regression Model is when we use one Independent Variable to predict values of Dependent variable. Ex: $Y_i = b_o + b_1X_i + e_i$.
- Multiple Regression Model is when we use more than one Independent Variable to predict the values of Dependent variable.
- ✓ A Regression Model with two Independent Variables (X_1 & X_2): $Y_i = b_o + b_1X_{1i} + b_1X_{2i} + e_i$.
- Regression Model with 3 Regressors ($X_{1,}$ X_2, & X_3): $Y_i = b_o + b_1X_{1i} + b_2X_{2i} + b_3X_{3i} + e_i$.
- Regression Model with N Regressors (X_1, X_2, …, X_n): $Y_i = b_o + b_1X_{1i} + b_2X_{2i} + … + b_nX_{ni} + e_i$.

Assessing Impact of Independent Variables

- Formula for hypothesis testing for significance of individual independent variables in the Regression model using a t-test is:

$$test\ statistic = \frac{\hat{\beta}_i - \beta_i^*}{SE\left(\hat{\beta}_i\right)}$$

- If the test is $H_0 : \beta_i = 0$

$$H_1 : \beta_i \neq 0$$

i.e., a test that the population coefficient is zero against a two-sided alternative, this is known as a *t*-ratio test:

Since $\beta_i{}^* = 0$, $test\ stat = \dfrac{\hat{\beta}_i}{SE(\hat{\beta}_i)}$

- Ratio of Coefficient to its SE is the *t*-ratio or *t*-statistic.
- To make appropriate inference regarding the impact of Independent variable on Dependent variable, we should assess the Significance, Sign and Size of its coefficient.

Assessing Goodness of Fit of Regression Model

- **R:** Correlation between Actual and Predicted values of Dependent Variable. Value close to 1, is better.
- **R Square:** Proportion of Variance in Dependent Variable which is explained by the Regression Model. Value closer to 1 (100%), is better.
- **Adjusted R Square:** Modification of R Square which takes into account the loss of degrees of freedom associated with adding extra variables. Should be close to 1.
- **F Statistic & Significance of F Statistic:** Tests the Null Hypothesis that Model has no explanatory power (R Square = 0) or Jointly tests the significance of all Regressors/Independent variables used in the regression model. Higher the F Statistic value and smaller the significance level, the better.


The end!
Thank you for viewing and listening!


HAVE A NICE DAY!