ONLINE GLOBAL WORKSHOP
ON
RESEARCH METHODOLOGY

MULTIPLE REGRESSION

Mukund Madhav Tripathi

mukund.m.tripathi@gmail.com

Concepts Need to KNOW

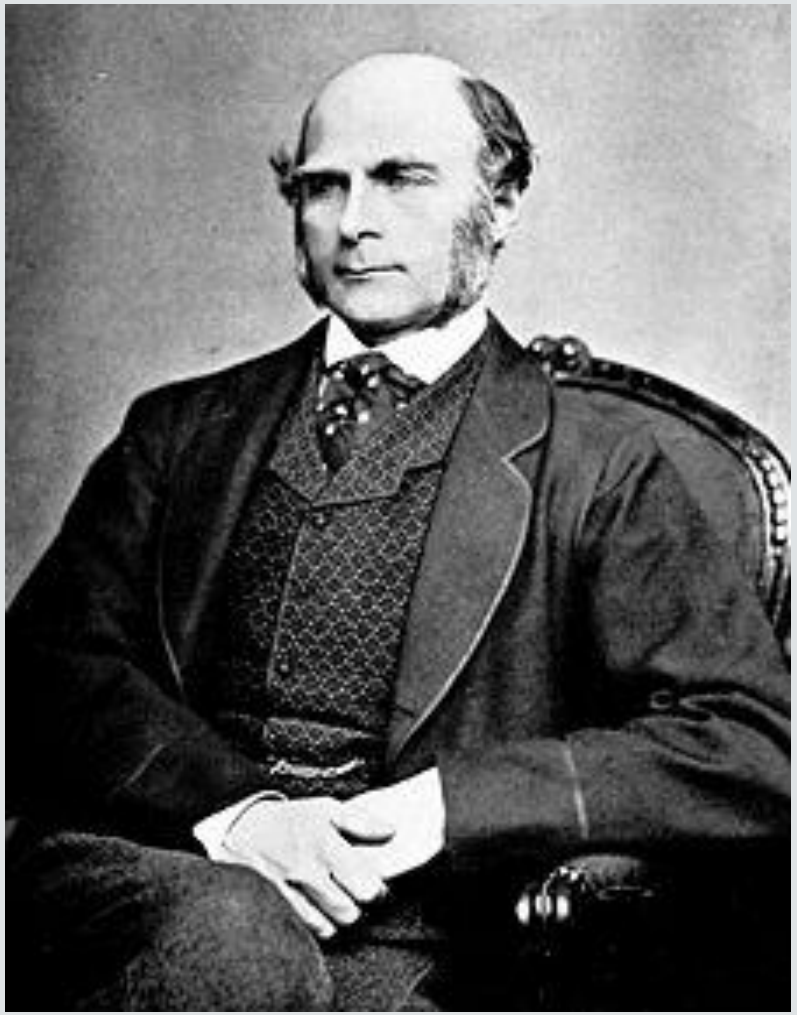
- Simple Linear Regression (SLR)
- Hypothesis and testing of hypothesis
- Coefficient of Determination
- Degree of freedom
- Correlation
- Variance
- Sample and sample testing
- Basics of Statistics
  - *Mean, Stand. Deviation, Frequency distribution, Normal distribution etc.*

Whom this session will benefit ?

- ✓ 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.

Regression History

Francis Galton



Francis Galton was the first to apply regression.

Claimed that height of children of tall parents "regress towards mean of that generation".

Modern regression analysis is developed by R A Fisher.

- Ref: F Galton, "Regression towards mediocrity in hereditary stature", *Nature*, Vol. 15, 246-263, 1886

Regression.

- An important tool in **Predictive Analytics**
- **Regression model establishes existence of association between two variables, but not causation.**

e.g.

Hypothesis - Married Men Earn More Money!

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

Regression is not designed to capture causal relationship

- **We use terms like Dependent and Independent**

So...
To Define Regression-

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.

It 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.

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

Frequent nomenclatures that are used for variables.

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

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

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

Multiple Linear Regression

Multiple linear regression means linear in regression parameters (beta values). The following are examples of multiple linear regression:.

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon$$

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \beta_4 x_2^2 \ldots + \beta_k x_k + \varepsilon$$

An important task in multiple regression is to estimate the beta values ($\beta_1, \beta_2, \beta_3$ etc...)

Regression: Matrix Representation

$$\begin{bmatrix} y_1 \\ y_2 \\ \\ \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{21} & x_{k1} \\ 1 & x_{12} & x_{22} & x_{k2} \\ & & & \\ & & & \\ 1 & x_{1n} & x_{2n} & x_{kn} \end{bmatrix} \bullet \begin{bmatrix} \beta_0 \\ \beta_1 \\ \\ \\ \beta_k \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \\ \\ \varepsilon_n \end{bmatrix}$$

$$Y = X\beta + \varepsilon$$

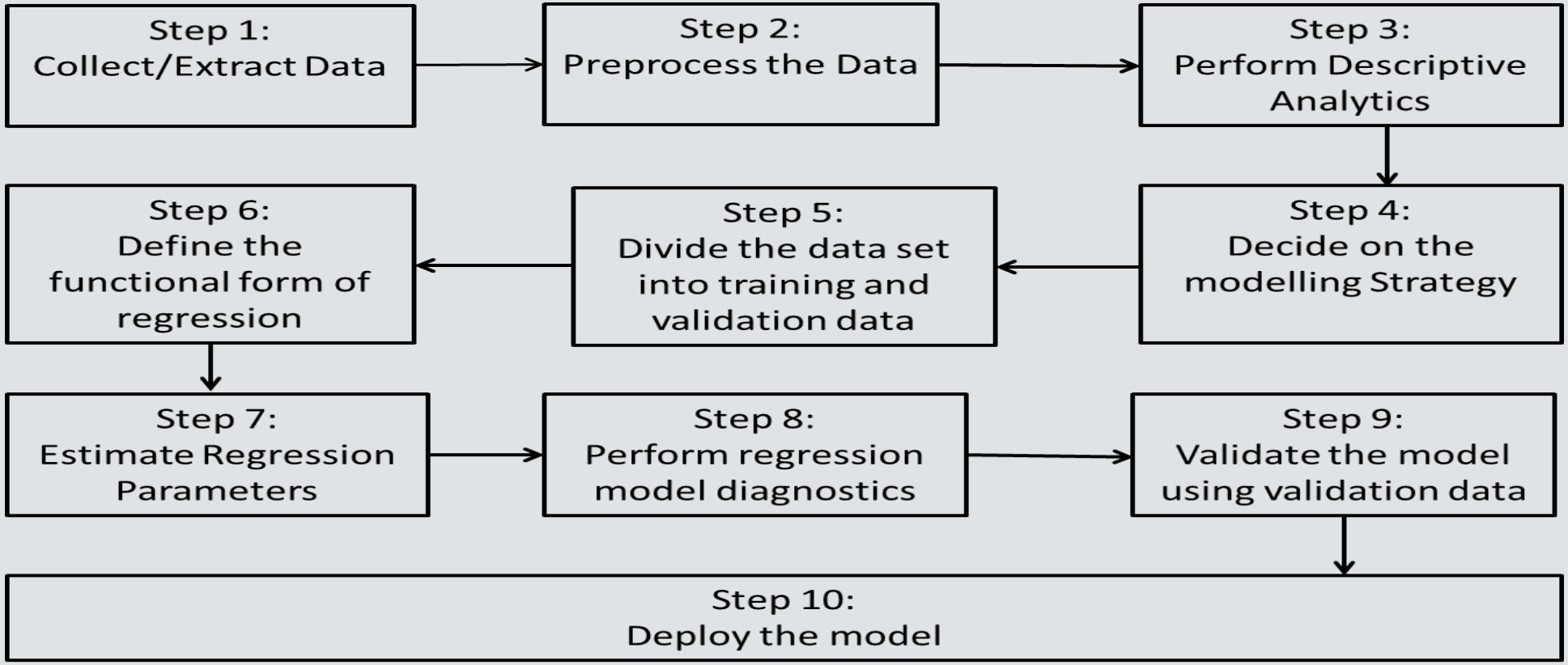
Multiple Linear Regression – Some Examples

- Estimating Travelling cost of transportation problems in linear programming problems.
- The treatment cost of a cardiac patient may depend on factors such as age, past medical history, body weight, blood pressure, and so on.
- Salary of MBA students at the time of graduation may depend on factors such as their academic performance, prior work experience, communication skills, and so on.
- Market share of a brand may depend on factors such as price, promotion expenses, competitors' price, etc.

ABC trucking company is an independent company in Gorakhpur district of U.P. A major portion of ABC's business involves deliveries of sugarcane throughout its local area. To develop better work schedules, the manager wants to predict the total daily travel time for their drivers

Manager believes that the total travel time would be closely related to total distance (in km) traveled in making the daily deliveries. He has taken a simple random samples of 10 driving assignments.
Sample data is given :

Framework for building multiple linear regression (MLR).



Collect/Extract Data

- The first step in regression model is to collect and/or extract data for the problem identified.

- In case of multiple linear regression model, we will have several independent variables.

Pre-process the Data

1) **Data Quality (measured through several characteristics such as completeness, correctness, etc.):** Data completeness refers to availability of necessary data for developing the model.
2) **Missing Data:** Many variables may have missing values. The data scientist has to come up with a strategy to handle missing values such as data imputation and specific techniques to carry out the imputation.
3) **Handling Qualitative Variables :** Qualitative variables or categorical variables need to be converted using dummy variables before incorporating them in regression model.
4) **Derive new variables** (such as ratios and interaction variables), which may have better association relationship with the dependent variable.

Perform Descriptive Analytics

- It is a good practice to start the MLR model building with descriptive analytics. In addition to descriptive statistics and data visualizations such as scatter plot and box plot.
- It is useful to check correlation between different variables since it can provide early warning for issues such as multi-collinearity.

Modelling Strategy

- When the number of variables runs into several hundreds, building regression models can get complicated due to multi-collinearity as well as computational complexity since estimation of regression parameters involves matrix inversion (Hat Matrix).

- The data scientist may also use specific variable selection approaches such as Forward Selection, Backward Elimination or Stepwise Regression

Estimate Regression Parameters

- Once the functional form is specified, the next step is to estimate the partial regression coefficients using the method of **Ordinary Least Squares** (OLS).
- OLS is used to fit a polygon through a set of data points, such that the sum of the squared distances between the actual observations in the sample and the regression equation is minimized.
- OLS provides the **Best Linear Unbiased Estimate** (BLUE).

Perform Regression Model Diagnostics

- F-test is used for checking the overall significance of the model whereas t-tests are used to check the significance of the individual variables. Presence of multi-collinearity can be checked through measures such as **Variance Inflation Factor** (VIF).

Validate the Model using Validation Data

The measures that can be used for validating the model in the validation data are as follows:

- R^2 or Adjusted R^2
- Root Mean Square Error (RMSE),

The final step in the regression model is to generate actionable items and the implementation plan.

Part (Semi-Partial) Correlation and Regression Model Building

The increase in the coefficient of determination, R^2, when a new variable is added is given by the square of the semi-partial correlation of the newly added variable with dependent variable Y.

Consider a regression model with two independent variables (say X_1 and X_2). The model can be written as follows:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon_i$$

Partial Correlation

Partial correlation is the correlation between the response variable Y and the explanatory variable X_1 when influence of X_2 is removed from both Y and X_1 (in other words, when X_2 is kept constant).

Alternatively, partial correlation is the correlation between residualized response and residualized explanatory variables.

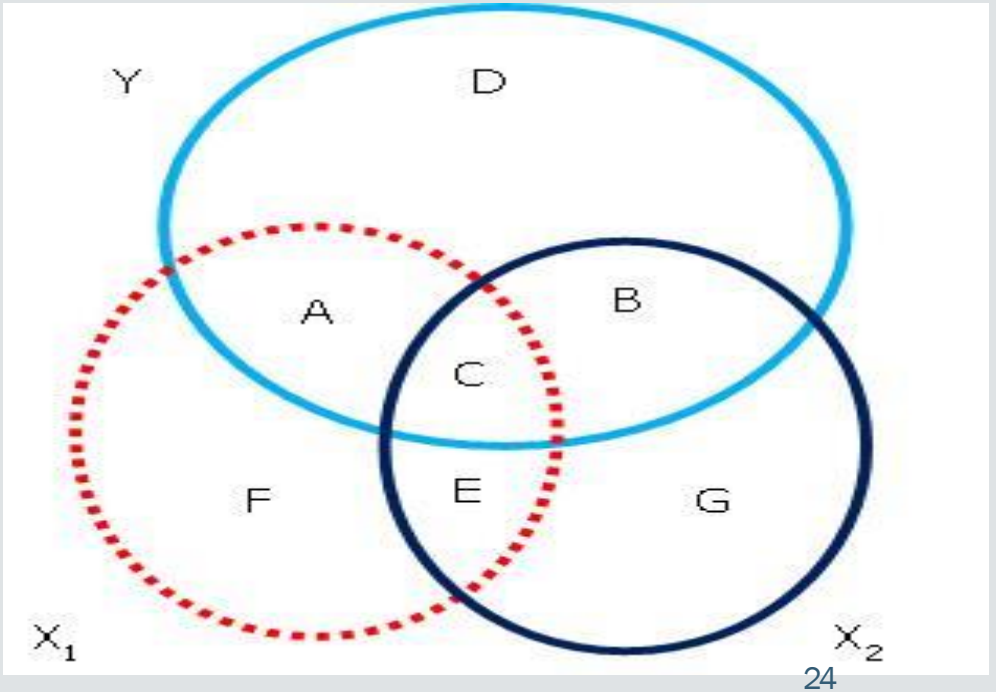
Let r_{YX_1,X_2} denote the partial correlation between Y and X_1 when X_2 is kept constant. Then r_{YX_1,X_2} is given by

$$r_{YX_1,X_2} = \frac{r_{YX_1} - r_{YX_2} \times r_{X_1X_2}}{\sqrt{(1 - r_{YX_2}^2) \times (1 - r_{X_1X_2}^2)}}$$

Semi-Partial Correlation (or Part Correlation)

- Consider a regression model between a response variable Y and two independent variables X_1 and X_2.
- The semi-partial (or part correlation) between a response variable Y and independent variable X_1 measures the relationship between Y and X_1 when the influence of X_2 is removed from only X_1 but not from Y.
- It is equivalent to removing portions C and E from X_1 in the Venn diagram shown in Figure
- Semi-partial correlation between Y and X_1, when influence of X2 is removed from X1 is given by

$$sr_{YX_1,X_2} = \frac{r_{YX_1} - r_{YX_1} r_{YX_2}}{\sqrt{(1 - r^2_{X_1X_2})}}$$



Semi-partial (part) correlation plays an important role in regression model building.

The increase in R-square (coefficient of determination), when a new variable is added into the model, is given by the square of the semi-partial correlation.

Co-efficient of Multiple Determination (*R*-Square) and Adjusted *R*-Square

As in the case of simple linear regression, *R*-square measures the proportion of variation in the dependent variable explained by the model. The co-efficient of multiple determination (*R*-Square or R^2) is given by

$$R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{\sum_{i=1}^{n} (\hat{Y}_i - Y_i)^2}{\sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2}$$

SSE is the sum of squares of errors and *SST* is the sum of squares of total deviation. In case of MLR, *SSE* will decrease as the number of explanatory variables increases, and *SST* remains constant.

To counter this, R2 value is adjusted by normalizing both SSE and SST with the corresponding degrees of freedom. The adjusted R-square is given below.

$$\text{Adjusted R - Square} = 1 - \frac{\text{SSE}/(n - k - 1)}{\text{SST}/(n - 1)}$$

THE CUMULATIVE TELEVISION RATING POINTS (*CTRP*) OF A TELEVISION PROGRAM, MONEY SPENT ON PROMOTION (DENOTED AS *P*), AND THE ADVERTISEMENT REVENUE (IN INDIAN RUPEES DENOTED AS *R*) GENERATED OVER ONE-MONTH PERIOD FOR 38 DIFFERENT TELEVISION PROGRAMS IS PROVIDED IN SHEET (TV)

Develop a multiple regression model to understand the relationship between the advertisement revenue (R) generated as response variable and promotions (*P*) and *CTRP* as

THANK YOU