DATA ANALYTICS

Unit 2:Linear Regression

Prof. Mamatha.H.R and Prof. Bharathi.R
Department of Computer Science and Engineering

DATA ANALYTICS

Unit 2:Linear Regression

Mamatha H R

Department of Computer Science and Engineering

Linear Regression

- Linear regression stands for a function that is linear in regression coefficients.

- The following equation will be treated as linear as far as regression is concerned.

$$Y = \beta_1 + \beta_1 X_1 + \beta_2 X_1 X_2 + \beta_3 X_2^2$$

Simple Linear Regression Model Building

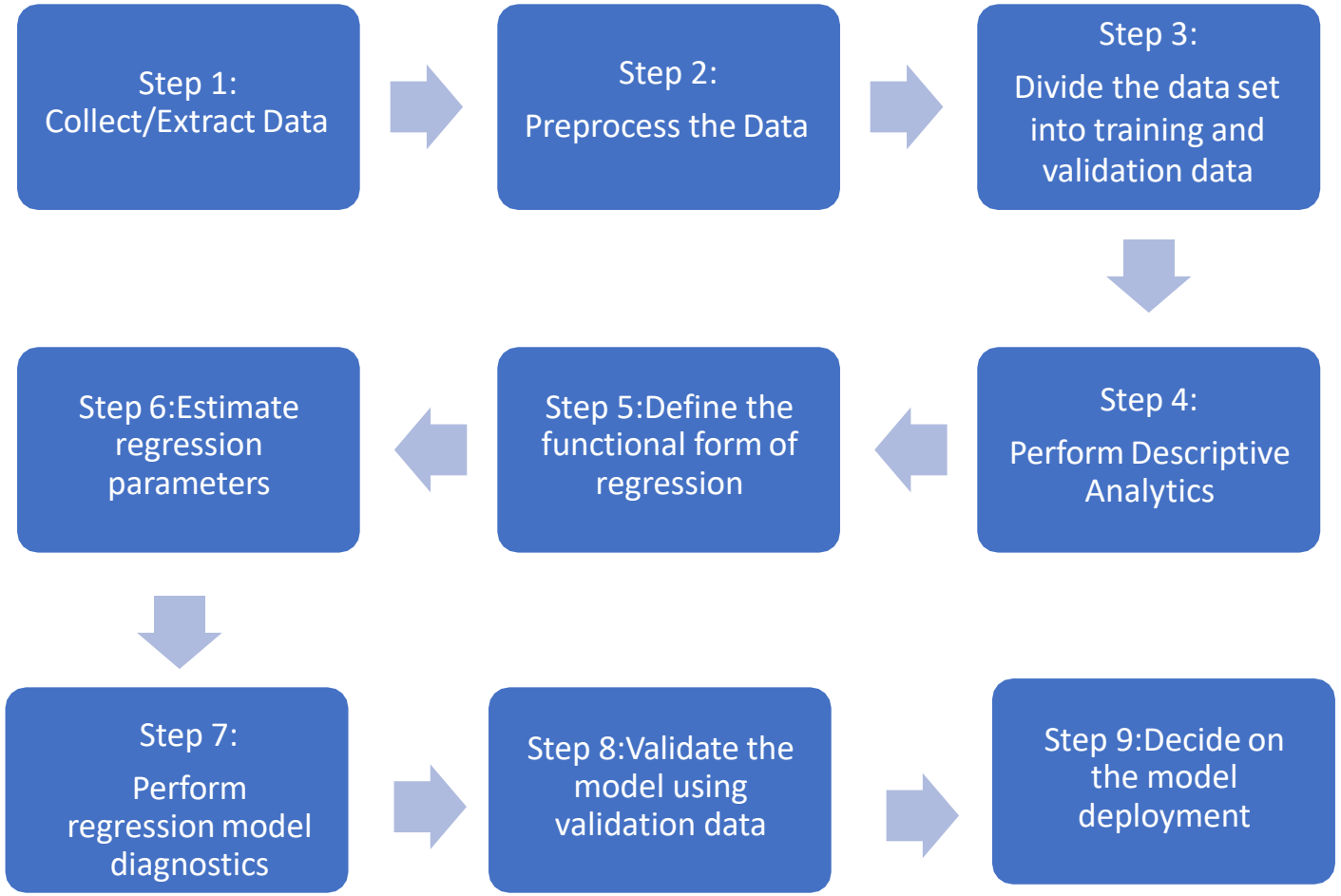
A simple linear regression model is developed to understand how the value of a KPI is associated with changes in the values of an independent variable.

Some examples are as follows:

1. A hospital may be interested in finding how the total treatment cost of a patient varies with the body weight of the patient.
2. E-commerce companies such as Amazon, Bigbasket and Flipkart would like to understand the number of customer visits to their portal and the revenue.
3. Retailers such as Walmart, Target, Reliance Retail, Hyper City, etc. would be interested in understanding the impact of price cut promotions on the revenue of their private labels (store brands or house brands).
4. Original equipment manufacturers (OEMs) would like to know the impact of duration of warranty on the profit.

DATA ANALYTICS

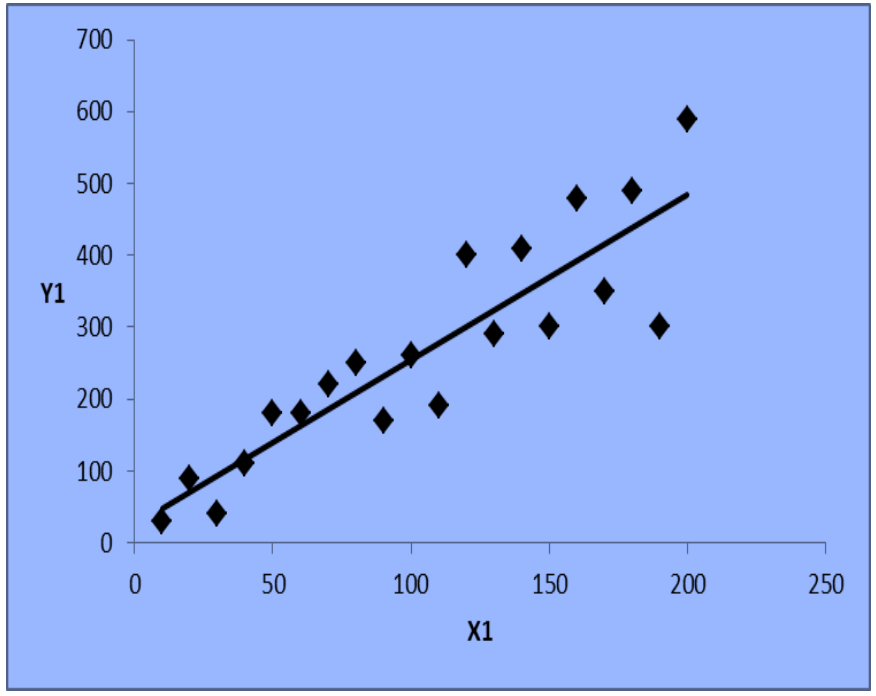
Framework for SLR model development



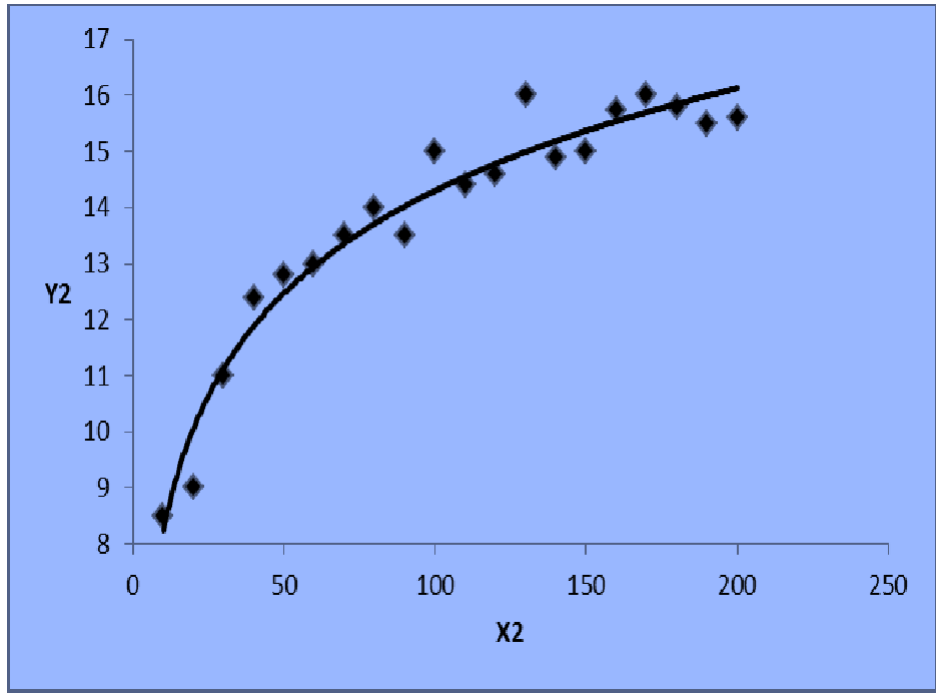
Define the Functional Form of Relationship

For better predictive ability (model accuracy) it is important to specify the correct functional form between the dependent variable and the independent variable. Scatter plots may assist the modeller to define the right functional form.

Linear relationship between X_1 and Y_1



Log-linear relationship between X_2 and Y_2.



Estimate the Regression Parameters

Once the functional form is specified, the next step is to estimate the regression parameters. The method of **Ordinary Least Squares (OLS)** is used to estimate the regression parameters.

OLS fits regression line through a set of data points such that the sum of the squared distances between the actual observations in the sample and the regression line is minimized (i.e., is minimized).

OLS provides the Best Linear Unbiased Estimate (BLUE). That is, where is the population parameter and is estimated parameter value from the sample.

Perform Regression Model Diagnostics

Regression is often misused since many times the modeller fails to perform necessary diagnostics tests before applying the model.

Before it can be applied it is necessary that the model created is validated for all model assumptions including the definition of the function form.

If the model assumptions are violated, then the modeller has to use some remedial measure; it is also possible that there is no association relationship between the variables at all.

Validate the Model using the Validation Data Set

A major concern in analytics is over-fitting, that is, the model may perform very well in the training data set but may perform badly in validation data set. It is important to ensure that the model performance is consistent in the validation data set as was in the training data set. In fact, the model may be cross-validated using multiple training and test data sets.

Decide on the Model Deployment

The final step in the regression model is to generate actionable items and business rules that can be used by the organization

Estimation of Parameters using Ordinary Least Squares

Given a set of dependent variable values (Y_i) and the corresponding independent variable values (X_i), each subject to a random error (ε_i), one has to find the best equation to represent the relationship between the dependent and independent variables.

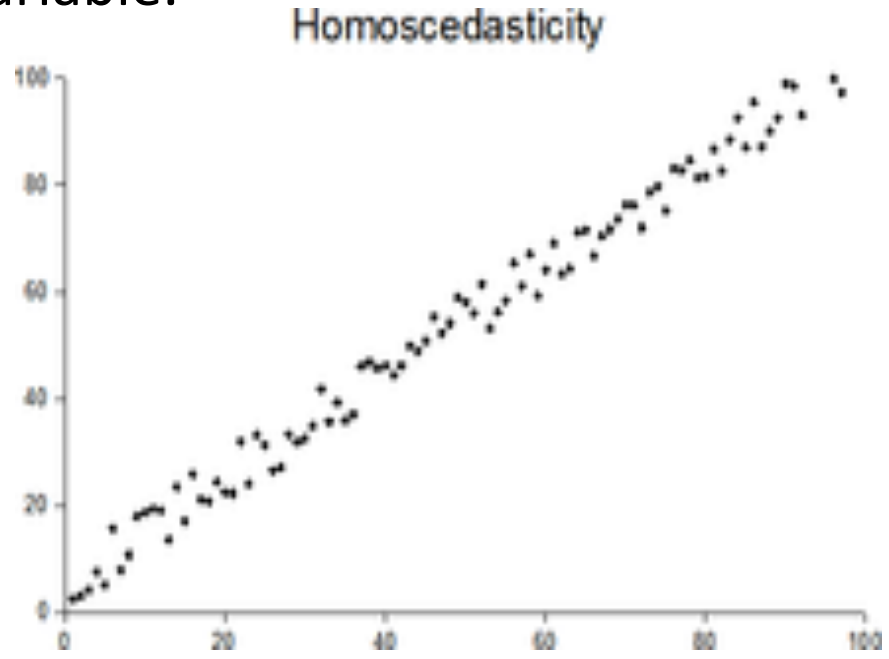
Assumptions

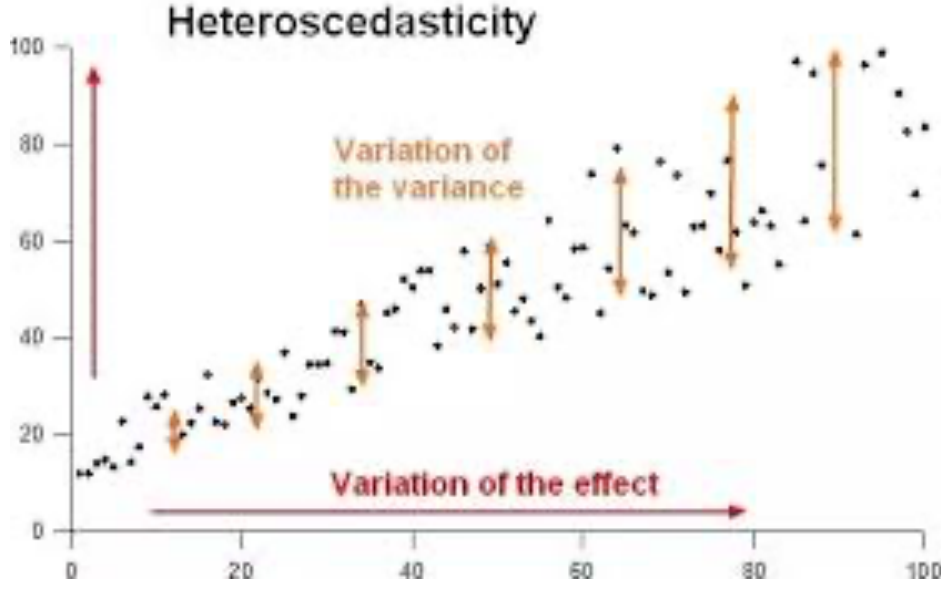
The method of least squares gives the best equation under the assumptions stated below (Harter 1974, 1975):

1. The regression model is linear in regression parameters.
2. The explanatory variable, X, is assumed to be non-stochastic (i.e., X is deterministic).
3. The conditional expected value of the residuals, $E(\varepsilon_i | X_i)$, is zero.
4. In case of time series data, residuals are uncorrelated, that is, $\text{Cov}(\varepsilon_i, \varepsilon_j) = 0$ for all $i \neq j$.
5. The residuals, ε_i, follow a normal distribution.
6. The variance of the residuals, Var(εi|Xi), is constant for all values of Xi. When the variance of the residuals is constant for different values of Xi, it is called homoscedasticity. A non- constant variance of residuals is called heteroscedasticity

Assumptions

The assumption of homoscedasticity (meaning “same variance”) is central to linear regression models. Homoscedasticity describes a situation in which the error term (that is, the “noise” or random disturbance in the relationship between the independent variables and the dependent variable) is the same across all values of the independent variables. Heteroscedasticity (the violation of homoscedasticity) is present when the size of the error term differs across values of an independent variable.





OLS Estimation

In ordinary least squares, the objective is find the optimal values of β_0 and β_1 that will minimize the **Sum of Squared Errors** (SSE) given in below Eq:

$$SSE = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$$

To find the optimal values of β_0 and β_1 that will minimize SSE, we have to equate the partial derivative of SSE with respect to β_0 and β_1 to zero.

$$\frac{\partial SSE}{\partial \beta_0} = \sum_{i=1}^{n} -2(Y_i - \beta_0 - \beta_1 X_i) = 2\left(n\beta_0 + \beta_1 \sum_{i=1}^{n} X_i - \sum_{i=1}^{n} Y_i\right) = 0$$

$$\frac{\partial SSE}{\partial \beta_1} = \sum_{i=1}^{n} -2X_i(Y_i - \beta_0 - \beta_1 X_i) = -2\sum_{i=1}^{n} (X_i Y_i - \beta_0 X_i - \beta_1 X_i^2) = 0$$

OLS Estimation

Solving the system of equations for β0 and β1, we get the estimated values as follows:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{X}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}\left(X_i Y_i - X_i \bar{Y}\right)}{\sum_{i=1}^{n}\left(X_i^2 - X_i \bar{X}\right)} = \frac{\sum_{i=1}^{n} X_i (Y_i - \bar{Y})}{\sum_{i=1}^{n} X_i (X_i - \bar{X})}$$

Example :

Salary of Graduating MBA Students versus Their Percentage Marks in Grade 10

Table in next slide provides the salary of 50 graduating MBA students of a Business School in 2016 and their corresponding percentage marks in grade 10 . Develop a linear regression model by estimating the model parameters.

DATA ANALYTICS

Salary of MBA students versus their grade 10 marks

S. No.	Percentage in Grade 10	Salary	S. No.	Percentage in Grade 10	Salary
1	62	270000	26	64.6	250000
2	76.33	200000	27	50	180000
3	72	240000	28	74	218000
4	60	250000	29	58	360000
5	61	180000	30	67	150000
6	55	300000	31	75	250000
7	70	260000	32	60	200000
8	68	235000	33	55	300000
9	82.8	425000	34	78	330000
10	59	240000	35	50.08	265000
11	58	250000	36	56	340000
12	60	180000	37	68	177600
13	66	428000	38	52	236000
14	83	450000	39	54	265000
15	68	300000	40	52	200000
16	37.33	240000	41	76	393000
17	79	252000	42	64.8	360000
18	68.4	280000	43	74.4	300000
19	70	231000	44	74.5	250000
20	59	224000	45	73.5	360000
21	63	120000	46	57.58	180000
22	50	260000	47	68	180000
23	69	300000	48	69	270000
24	52	120000	49	66	240000
25	49	120000	50	60.8	300000

Solution

Using Eqs., the estimated values of β_0 and β_1 are given by

$$\hat{\beta}_0 = 61555.3553 \text{ and } \hat{\beta}_1 = 3076.1774$$

The corresponding regression equation is given by

$$\hat{Y}_i = 61555.3553 + 3076.1774 X_i$$

Where $\hat{Y}_i$ is the predicted value of Y for a given value of X_i. .

The equation can be interpreted as follows:

for every one percentage increase in grade 10 marks, the salary of the MBA students will increase at the rate of 3076.1774 on an average. The notations

Solution Continued

$\hat{\beta}_0$ and $\hat{\beta}_1$ are used to denote that these are estimated values of the regression coefficients from the sample of 50 students.

Regression coefficient estimates using Microsoft Excel

	Coefficients	Standard Error	t-stat	p-value
Intercept	61555.35534	66701.901	0.9228	0.3607
Percentage in grade 10	3076.177438	1031.5258	2.9821	0.0044

Interpretation of Simple Linear Regression Coefficients

- Interpretation of regression coefficients is important for understanding the relationship between the response variable and the explanatory variable and the impact of change in the values of explanatory variables on the response variable.

- The interpretation will depend on the functional form of the relationship between the response and the explanatory variables.

DATA ANALYTICS

Interpretation of β_0 and β_1 in $Y = \beta_0 + \beta_1 X$

When the functional form is $Y = \beta_0 + \beta_1 X$,

the value of $\beta_0 = E(Y/X=0)$.

$\beta_1 = \frac{\partial Y}{\partial X}$ that is β_1 is the change in the value of Y for the unit change in the value of X.

Where $\frac{\partial Y}{\partial X}$ is the partial derivative of Y with respect to X.

Exercise

A car dealership records the age of cars(in years) and price for cars(Rs) it sold in the last year

Car Age in yrs	8	8	9	9	10	10	11	12	13
Price in Rs	450000	400000	320000	310000	250000	210000	260000	240000	220000
Car Age in yrs	4	4	4	5	5	6	7	7	8
Price in Rs	620000	570000	680000	560000	450000	490000	460000	430000	420000

Using the data above, answer the following questions.

a. If a car is seven years old, what price could we expect?

b. Does β_0 have a meaningful interpretation?

Validation of the Simple Linear Regression Model

It is important to validate the regression model to ensure its validity and goodness of fit before it can be used for practical applications. The following measures are used to validate the simple linear regression models:

1. Co-efficient of determination (*R*-square).
2. Hypothesis test for the regression coefficient β_1
3. Analysis of Variance for overall model validity (relevant more for multiple linear regression).
4. Residual analysis to validate the regression model assumptions.
5. Outlier analysis.

- The above measures and tests are essential, but not exhaustive.

DATA ANALYTICS

Exercise

1. Explain with example "Regression establishes existence of an association relationship between two variables, and not a causal relationship".
2. Reading Exercise, "Rubin Casual Model".

❑

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