LECTURE - 45

↳ We will standardise the given input data and then we start the step by step procedure of linear Regression i.e.,

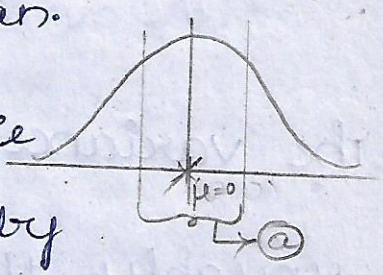
→ Split the train test

→ Train the model i.e., X_train, y_train.

→ y_train_predicted, y_train → ERROR ON TRAIN DATA

* The residuals should be normally distributed with zero mean.

@ ⇒ Most of the values lie near to the mean - by looking at the Residuals (ERROR's)



* The residuals should be independent of each other.

* Homoscedasticity - The variance should be constant at the residuals on the training data.

→ Prediction on X_test $\xrightarrow{O/P}$ Y test_predicted.

→ Evaluation on Y_test, Y_test_predicted by means of MSE; MAE, R^2-SCORE, RMSE.

HOMOSCEDASTIC :

It refers to a condition in which the variance of the residual or the error term in a regression model is constant

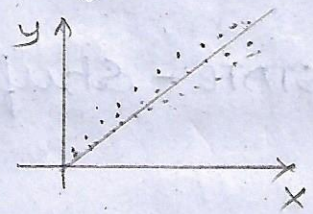
(or)

The variance of the data points are roughly the same for all the data points.

ASSUMPTIONS OF LINEAR REGRESSION:

1. LINEAR RELATIONSHIP:

There should exists a linear relationship between the independent (x or INPUT) and the dependent (y or OUTPUT) variables.



2. INDEPENDENCE:

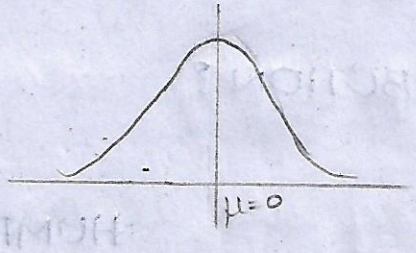
The residuals are independent.

3. HOMOSCEDASTICITY:

The residuals have constant variance at every level of X.

4. NORMALITY:

The residuals of the model are normally distributed.



HOW DO WE CHECK THE LINEAR REGRESSION MODEL?

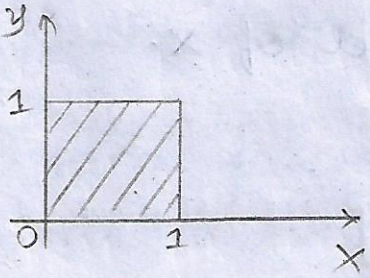
→ By understanding and visualizing the data

i.e., SCATTER PLOT.

↳ RANDOM STATE - Shuffle should be constant

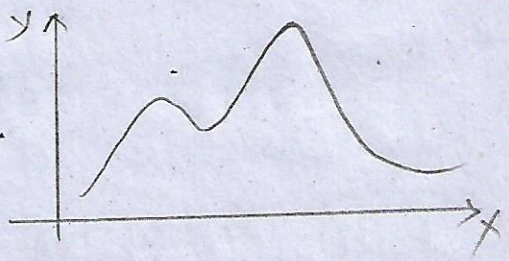
↳ NORMALISATION - Also called as "MIN-MAX SCALER".

* It removes the scale dependency.
* The values range between '0' & '1'.



WHY DO WE HAVE HUMPS IN NORMAL DISTRIBUTION?

HUMPS →



↳ We will have humps because of the limited data.

↳ If the data is large/have more values, then there will be no humps in the distribution and we get a perfect normal distribution.

MULTIPLE LINEAR REGRESSION:

A regression model that estimates the relationship between a quantitative dependent variable and two (or) more independent variables using a straight line.

80-20 SPLIT:

The 80-20 rule, also known as pareto principle, is an aphorism which asserts that 80% of outcomes (or) OUTPUTS result from 20% of all causes (or) INPUTS for any given event.

→ In business, a goal of the 80-20 rule is to identify inputs that are potentially the most productive and make them the priority.

WHY DO WE SPLIT THE DATASET INTO TRAINING & TEST DATA?

↳ Seperating data into training and testing sets is an important part of evaluating data mining models.

↳ By using similar data for training and testing, we can minimize the effects of data discrepancies and better understand the characteristics of the model.

WHEN SHOULD WE SPLIT DATA?

↳ If we have less training data, the parameters estimates have greater variance.

↳ If we have less testing data, the performance statistic will have a greater variance.

↳ the data should be divided in such a way that neither of them is too high, which is more dependent on the amount of data we have.