SNS COLLEGE OF TECHNOLOGY

An Autonomous Institution
Coimbatore-35

DEPARTMENT OF ARTIFICIAL INTELLIGENCE & DATA SCIENCE

23ADT202 – FUNDAMENTALS OF DATA SCIENCE AND ANALYTICS

II YEAR IV SEM

UNIT II – Interpretation of R^2

Interpretation of R^2

EMPATHY:

- 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 b1.

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 primary objective of regression is to explain the variation in Y using the knowledge of X.
- The coefficient of determination (R-square) measures the percentage of variation in Y explained by the model (ß0 + ß1 X).

DEFINE:

Characteristics of R-square:

- Here are some basic characteristics of the measure:

1. Since R^2 is a proportion, it is always a number between 0 and 1.

2. If $R^2 = 1$, all of the data points fall perfectly on the regression line. The predictor x accounts for all of the variation in y!.

3. If $R^2 = 0$, the estimated regression line is perfectly horizontal. The predictor x accounts for none of the variation in y!

IDEATE:

- Coefficient of determination, R^2 a measure that assesses the ability of a model to predict or explain an outcome in the linear regression setting.

- More specifically, R2 indicates the proportion of the variance in the dependent variable (Y) that is predicted or explained by linear regression and the predictor variable (X, also known as the independent variable).

PROTOTYPING:

- In general, a high R^2 value indicates that the model is a good fit for the data, although interpretations of fit depend on the context of analysis.

- An R^2 of 0.35, for example, indicates that 35 percent of the variation in the outcome has been explained just by predicting the outcome using the covariates included in the model.

- That percentage might be a very high portion of variation to predict in a field such as the social sciences; in other fields, such as the physical sciences, one would expect R^2 to be much closer to 100 percent.

- The theoretical minimum R2 is 0.

- However, since linear regression is based on the best possible fit, R2 will always be greater than zero, even when the predictor and outcome variables bear no relationship to one another.

TESTING:

- R^2 increases when a new predictor variable is added to the model, even if the new predictor is not associated with the outcome.

- To account for that effect, the adjusted R^2 incorporates the same information as the usual R^2 but then also penalizes for the number of predictor variables included in the model.

- As a result, R^2 increases as new predictors are added to a multiple linear regression model, but the adjusted R increases only if the increase in R^2 is greater than one would expect from chance alone.

- In such a model, the adjusted R^2 is the most realistic estimate of the proportion of the variation that is predicted by the covariates included in the model.