PUNE INSTITUTE OF COMPUTER TECHNOLOGY

Subject: Machine Learning (LP-1 LAB)

Name: Aditya Kangune **Roll No. : 33323**

Batch: K11 **Academic Year: 2021-22**

Assignment 2

Regression technique

Problem Statement:

This data consists of temperatures of INDIA averaging the temperatures of all places month-wise.

Temperatures values are recorded in CELSIUS

A. Apply Linear Regression using suitable library function and predict the Month-wise temperature.
B. Assess the performance of regression models using MSE, MAE, and R-Square metrics
C. Visualize a simple regression model.

Objective:

This assignment will help the students to realize how Linear Regression can be used and predictions using the same can be performed.

Theory:

Definition of Linear Regression:

1. In layman’s terms, we can define linear regression as it is used for learning the linear relationship between the target and one or more

forecasters, and it is probably one of the most popular and well-inferential algorithms in statistics.

2. Linear regression endeavors to demonstrate the connection between two variables by fitting a linear equation to observed information.
3. One variable is viewed as an explanatory variable, and the other is viewed as a dependent variable.

Types of Linear Regression:

Normally, linear regression is divided into two types: Multiple linear regression and Simple linear regression.

1. Multiple Linear Regression:

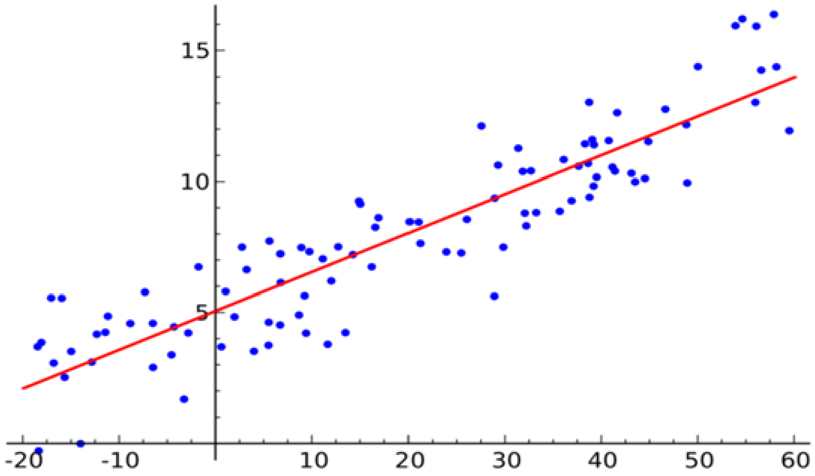
A. In this type of linear regression, we always attempt to discover the relationship between two or more independent variables or inputs and the corresponding dependent variable or output and the independent variables can be either continuous or categorical.

B. This linear regression analysis is very helpful in several ways like it helps in foreseeing trends, future values, and moreover predicting the impacts of changes.

2. Simple Linear Regression:

In simple linear regression, we aim to reveal the relationship between a single independent variable or you can say input, and a corresponding dependent variable or output. We can discuss this in a simple line as

$$y = \beta 0 + \beta 1x + \varepsilon$$



Simple Linear Regression graph [3]

Here, Y speaks to the output or dependent variable, β0 and β1 are two obscure constants that speak to the intercept and coefficient that is slope separately, and the error term is ε Epsilon.
We can also discuss this in the form of a graph and here is a sample simple linear regression model graph.

What Actually is Simple Linear Regression?

It can be described as a method of statistical analysis that can be used to study the relationship between two quantitative variables.

Primarily, there are two things that can be found out by using the method of simple linear regression:

1. Strength of the relationship between the given duo of variables:

For example, the relationship between global warming and the melting of glaciers

2. How much the value of the dependent variable is at a given value of the independent variable:
For example, the amount of melting of a glacier at a certain level of global warming or temperature

1. Regression models are used for the lab
2. orated explanation of the relationship between two given variables. There are certain types of regression models like logistic regression models, nonlinear regression models, and linear regression models.
3. The linear regression model fits a straight line into the summarized data to establish the relationship between two variables.

Assumptions of Linear Regression:

To conduct a simple linear regression, one has to make certain assumptions about the data. This is because it is a parametric test.

The assumptions used while performing a simple linear regression are as follows:

- **Homogeneity of variance (homoscedasticity):** One of the main predictions in a simple linear regression method is that the size of the error stays constant.

This simply means that in the value of the independent variable, the error size never changes significantly.

- **Independence of observations:** All the relationships between the observations are transparent, which means that nothing is hidden, and only valid sampling methods are used during the collection of data.

- **Normality:** There is a normal rate of flow in the data.

These three are the assumptions of regression methods.

However, there is one additional assumption that has to be taken into consideration while specifically conducting a linear regression.

- **The line is always a straight line:** There is no curve or grouping factor during the conduction of linear regression. There is a linear relationship between the variables (dependent variable and independent variable). If the data fails the assumptions of homoscedasticity or normality, a nonparametric test might be used. (For example, the Spearman rank test)

Example of data that fails to meet the assumptions:

One may think that cured meat consumption and the incidence of colorectal cancer in the U.S have a linear relationship.

But later on, it comes to the knowledge that there is a very high range difference between the collection of data of both the variables.

Since the homoscedasticity assumption is being violated here, there can be no linear regression test.

However, a Spearman rank test can be performed to know about the relationship between the given variables.

Applications of Simple Linear Regression:

1. **Marks scored by students based on the number of hours studied (ideally):**

Here marks scored in exams are dependent and the number of hours studied is independent.

2. **Predicting crop yields based on the amount of rainfall:**

Yield is a dependent variable while the measure of precipitation is an independent variable.

3. **Predicting the Salary of a person based on years of experience:**

Therefore, Experience becomes independent while Salary turns into the dependent variable.

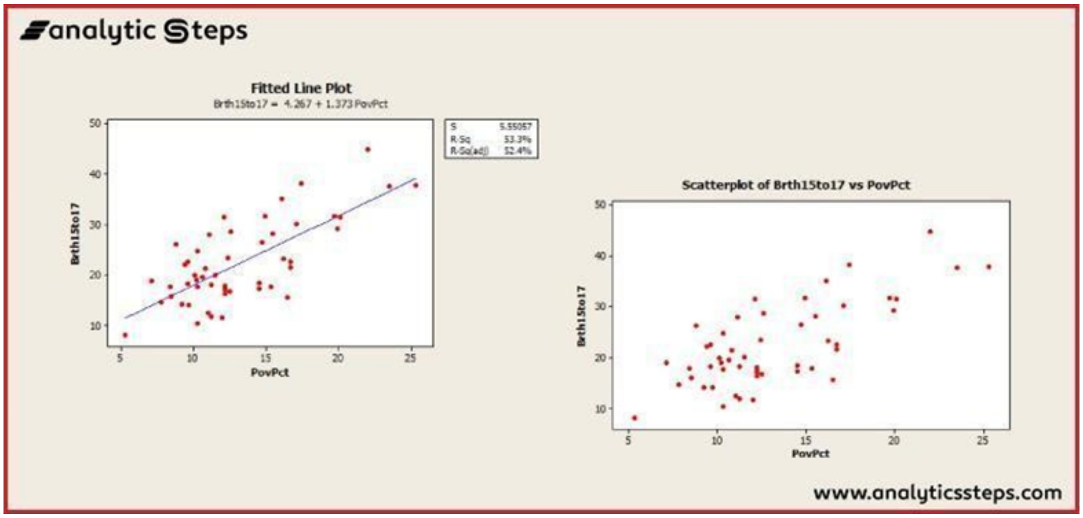
Limitations of Simple Linear Regression:

1. Indeed, even the best information doesn't recount a total story. Regression investigation is ordinarily utilized in examination to set up that a relationship exists between variables.
2. However, correlation isn't equivalent to causation: a connection between two variables doesn't mean one causes the other to occur.
3. Indeed, even a line in a simple linear regression that fits the information focuses well may not ensure circumstances and logical results relationship.
4. Utilizing a linear regression model will permit you to find whether a connection between variables exists by any means.
5. To see precisely what that relationship is and whether one variable causes another, you will require extra examination and statistical analysis.

Examples of Simple Linear Regression:

1. Now, let's move towards understanding simple linear regression with the help of an example.
2. We will take an example of teen birth rate and poverty level data.
3. This dataset of size n = 51 is for the 50 states and the District of Columbia in the United States. The variables are y = year 2002 birth rate for every 1000 females 15 to 17 years of age and x = destitution rate, which is the percent of the state's populace living in families with wages underneath the governmentally characterized neediness level. (Information source: Mind On Statistics, 3rd version, Utts and Heckard).
4. Below is the graph in which you can see the (birth rate on the vertical) is indicating a normally linear relationship, on average, with a positive slope.

5.



Example graph of simple linear regression [3]

6. As the poverty level builds, the birth rate for 15 to 17-year-old females will in general increment too.

Here is another graph which is showing a regression line superimposed on the data:

1. The condition of the fitted regression line is given close to the highest point of the plot. The condition should express that it is for the "average" birth rate (or "anticipated" birth rate would be alright as well) as a regression condition portrays the normal estimation of y as a component of at least one x-variables. In statistical documentation, the condition could be composed y^=4.267+1.373x.
2. The interpretation of the slope (value = 1.373) is that the 15 to 17-year-old birth rate increases 1.373 units, on average, for each one unit (one percent) increase in the poverty rate.
3. The translation of the intercept (value=4.267) is that if there were states with a population rate = 0, the anticipated normal for the 15 to 17-year-old birth rate would be 4.267 for those states.
4. Since there are no states with a poverty rate = 0 this understanding of the catch isn't basically significant for this model.
5. In the chart with a repression line present, we additionally observe the data that s = 5.55057 and r2 = 53.3%.

6. The estimation of s discloses to us generally the standard deviation of the contrasts between the y-estimations of individual perceptions and expectations of y dependent on the regression line.
7. The estimation of r2 can be deciphered to imply that destitution rates
8. "clarify" 53.3% of the noticed variety in the 15 to 17-year-old normal birth paces of the states.
9. The R2 (adj) value (52.4%) is a change in accordance with R2 dependent on the number of x- variables in the model (just one here) and the example size. With just a single x-variable, the charged R2 isn't significant.

Conclusion:

Simple linear regression is a regression model that figures out the relationship between one independent variable and one dependent variable using a straight line.

Implementation: