Assumptions Linear Regression

Linear Regression:

Linear regression is an analysis used to predict the relationship between two variables. A rule of thumb for the sample size is that regression analysis requires at least 20 cases per independent variable in the analysis.

In order to assume that our predicted relation is good, the linear regression should satisfy the following assumption which are listed below

Assumptions of Linear Regression:

1. **Linear relation between X and Y**
2. **Data should have less or no multicollinearity**
3. **Residual(variables) should be normally distributed**
4. **Residual(variables) should have Homoscedasticity**

For example, In an engineering class room, a teacher asked 20 people (by taking samples of people with high grade, people with average grade and people with more arrears) to take an assignment which is to note the date to day activities like studying hours, sleep time, playing time, using mobile phones etc.. then he has plotted the graph with related to these values, and he found that the below results with respect to linear regression.

Linear relation between X and Y

-> As the name Linear regression, says the linear relation between two variables, to satisfy that our independent variable (x) and dependent variable (y) should be correlated.

-> From the above example, it was found that persons who concentrated more on studies with less sleep time has more grade, and the person whose utilized the average amount has average grade, and the person who use less time has less grade. From this assumption it was clearly noted that there was a linear relationship between the studying time and marks obtained.

Assumptions Linear Regression

-> Also, we can find that some points are above or below the line of regression. These points that lie outside the line of regression are the outliers. It is also important to check for outliers since linear regression is sensitive to outlier effects

-> If the correlation is positive, regression slope is positive. If the correlation is Negative, regression slope is negative.

Residual(variables) should be normally distributed:

-> In a Linear regression all the variable should be normally distributed. This is the second thing for the linear regression analysis that requires all variables to be multivariate normal.

-> From the previous values that students reported their activities like studying, sleeping, and engaging in social media. Now, all these activities have a relationship with each other. If you study for a more extended period, you sleep for less time. Similarly, extended hours of study affect the time you engage in social media.

Data should have less or no multicollinearity:

-> Most critical assumption in linear regression is multi-collinearity, because it affects the statically inference between the variable. For example, the variable data has a relationship, but they do not have much collinearity.

-> From previous values, it was found that person who has utilized the minimum hours in study got good grade and the person who utilized more time in social media also got good grade. So, the point is that there is a relationship, but not a multicollinear one.

-> Also, during covid, they have conducted the online exam where most of the people got good grade, these are also the example of multi-collinearity.

-> If you still find some amount of multicollinearity in the data, the best solution is to remove that multicollinearity data in our model, which makes our assumption of linear regression a critical one.

Assumptions Linear Regression

Residual(variables) should have Homoscedasticity:

As the name Homoscedasticity [Homo – Same, scedasticity – variance] says that the variable should have the same or similar type of variance. So, this make the data across the line of regression is equal.