MULTICOLLINEARITY

Multicollinearity, also collinearity is a phenomenon in which one feature variable in a regression model is highly linearly correlated with another feature variable.

Multicollinearity is when two or more predictors in a regression are highly related to one another, such that they do not provide unique independent information to the regression. Thus It becomes difficult for the model to estimate the relationship between each independent variable and the dependent variable independently.

In other words if two independent variables are positively or negatively correlated to each other, it is called multicollinearity.

Multicollinearity makes it hard to interpret your coefficients or reduces the precision of the estimate coefficients and it reduces the power of your model to identify independent variables that are statistically significant.

It can be calculated using

- Correlation matrix

This matrix has cells where each cell shows correlation of one variable with one other variable. So this correlation value is calculated for one variable vs one other variable.

- Variance Inflation Factor (VIF)

VIF measures how much of the variation in one variable is explained by all the other variable. So this vif value is calculated for one variable vs all other variables

It is obtained by regressing each independent variable, say X on the remaining independent variables (say Y and Z) and checking how much of it (of X) is explained by these variables.

VIF score of a variable say X is given by

Vif = 1/(1-R2)

where R2 is the variance calculated by making a regression model using X as dependent variable and all others as independent variables.

VIF starts at 1 and has no upper limit

VIF = 1, no correlation between the independent variable and the other variables

VIF exceeding 5 starts indicating high multicollinearity between this independent variable and the others

Solution to Multicollinearity

Remove independent variables with high VIF values. Dropping variables should be an iterative process starting with the variable having the largest VIF value because its trend is highly captured by other variables. If you do this, you will notice that VIF values for other variables would have reduced too, although to a varying extent.