**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.

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

When computing the matrix of Pearson’s Bivariate Correlation among all independent variables the correlation coefficients need to be smaller than 1.

- Variance Inflation Factor (VIF)

VIF measures how much of the variation in one variable is explained by the other variable.

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 is given by

https://miro.medium.com/max/268/1*NeeF1ZiblNTcvdBbNjWSCA.png

VIF starts at 1 and has no upper limit

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

VIF exceeding 5 or 10 indicates high multicollinearity between this independent variable and the others

What causes Multicollinearity?

- Multicollinearity could exist because of the problems in the dataset at the time of creation.

- Multicollinearity could also occur when new variables are created which are dependent on other variables. For example, creating a variable for BMI from the height and weight variables would include redundant information in the model

- Including identical variables in the dataset. For example, including variables for temperature in Fahrenheit and temperature in Celsius

- Inaccurate use of dummy variables can also cause a multicollinearity problem. This is called the Dummy variable trap. For example, in a dataset containing the status of marriage variable with two variables ‘married’, ’single’. Creating dummy variables for both of them would include redundant information. We can make do with only one variable containing 0/1 for ‘married’/’single’ status.

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.