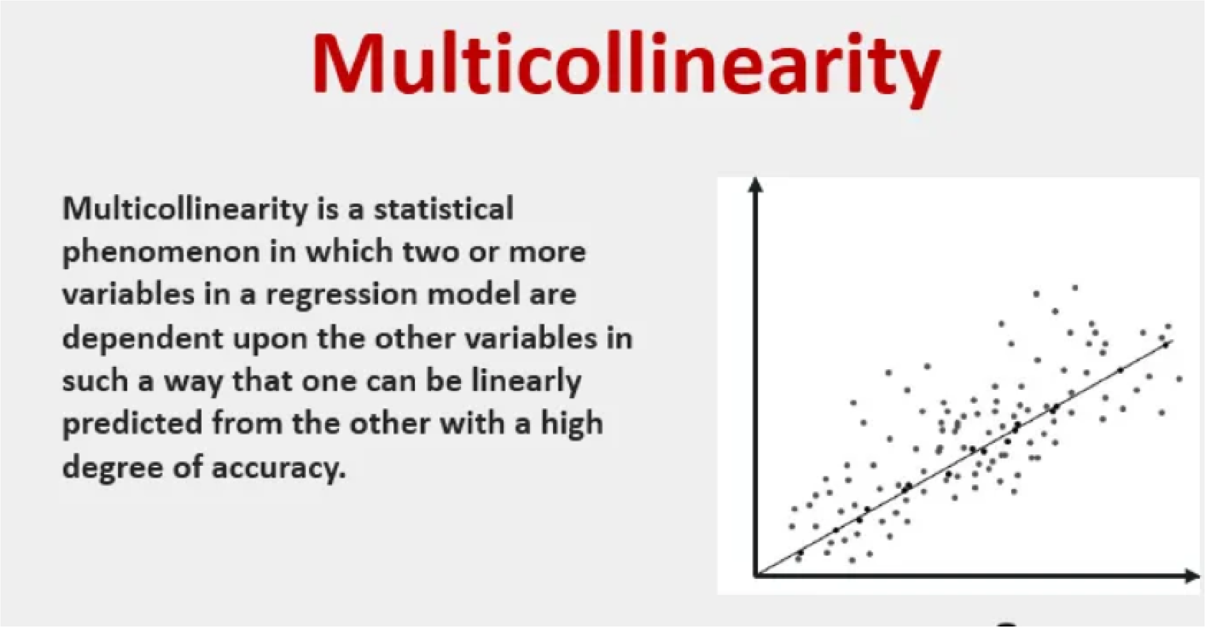

Multicollinearity

Multicollinearity is a statistical phenomenon in which two or more variables in a regression model are dependent upon the other variables in such a way that one can be linearly predicted from the other with a high degree of accuracy.

To avoid this problem, we can use statistical methods such as variance inflation factor (VIF) or correlation matrix to identify the variables with high correlation and remove one of the variables to reduce multicollinearity.

In summary, multicollinearity is a situation in which independent variables in a regression model are highly correlated with each other, making it difficult to determine the independent effect of each variable on the outcome variable. In the example of predicting salary using CGPA, IQ, and package, multicollinearity can arise if CGPA and IQ are highly correlated with each other and can be addressed by identifying and removing one of the correlated variables.

Types of Multicollinearities

There are two main types of multicollinearity:

1. Structural multicollinearity: This type of multicollinearity arises from the way variables are defined or the construction of the model. It occurs when one independent

variable can be expressed as a linear combination of other independent variables. For example, if variable A is equal to 2 times variable B plus 3 times variable C, then there is structural multicollinearity.

2. Data-driven multicollinearity: This type of multicollinearity occurs due to the specific data being analyzed. It arises when the independent variables in the dataset are highly correlated with each other, regardless of how they are defined or the model construction. It is a result of the observed patterns in the data.

Both types of multicollinearity can pose challenges in regression analysis, as they can lead to unstable or unreliable estimates of the regression coefficients and affect the interpretation of the model. Detecting and addressing multicollinearity is important to ensure the validity and accuracy of the regression results.

Variance Inflation Factor (VIF)

Variance Inflation Factor (VIF) is a metric used to quantify the severity of multicollinearity in a multiple linear regression model. It measures the extent to which the variance of an estimated regression coefficient is increased due to multicollinearity.

The VIF for the predictor variable is then calculated as the reciprocal of the variance explained by the other predictors, which is equal to 1 / (1 — R^2). Here, R^2 is the coefficient of determination for the linear regression using the predictor variable as the response variable.

The VIF calculation can be summarized in the following steps:

1. For each predictor variable X_i in the regression model, perform a linear regression using X_i as the response variable and the remaining predictor variables as the independent variables.
2. Calculate the R^2 value for each of these linear regressions.
3. Compute the VIF for each predictor variable X_i as $VIF_i = 1 / (1 — R^2_i)$

A VIF value close to 1 indicates that there is very little multicollinearity for the predictor variable, whereas a high VIF value (e.g., greater than 5 or 10, depending on the context) suggests that multicollinearity may be a problem for the predictor variable, and its estimated coefficient might be less reliable.

Keep in mind that VIF only provides an indication of the presence and severity of multicollinearity and does not directly address the issue. Depending on the VIF values and the goals of the analysis, you might consider using techniques like variable selection, regularization, or dimensionality reduction methods to address multicollinearity.

STEPS TO AVOID MULTICOLLINEARITY:

- Set VIF value and remove variables above the value.
- Use Regularization techniques like Ridge, Lasso and Elastic Net.
- Using heatmap/Correlation Matrix to detect the highly collinear variables and remove them manually
- Using Feature Engineering, Combine the correlated variables.