Methods to Find Multicollinearity

Correlation Matrix:

A simple method to visually identify high correlations between independent variables, where values close to 1 or -1 indicate potential multicollinearity.

- Values close to +1 or -1 indicate high multicollinearity.
- Values close to 0 indicate weak or no correlation.
- A rule of thumb: correlation > 0.75 suggests multicollinearity.

Variance Inflation Factor (VIF):

A quantitative measure that indicates how much the variance of a regression coefficient is inflated due to the presence of correlated variables; a high VIF value (typically above 10) suggests multicollinearity.

- Values above 10 are often considered a sign that the model's multicollinearity should be reduced.

Tolerance & VIF (Alternative Approach):

Tolerance is the inverse of VIF and is another way to check for multicollinearity.

- Tolerance < 0.1 → High multicollinearity.
- VIF > 10 → Severe multicollinearity.

Eigenvalue analysis:

The condition number detects multicollinearity by checking the ratio of the largest to the smallest eigenvalue of the feature matrix. Analysing the eigenvalues of the correlation matrix, where very small eigenvalues can indicate potential multicollinearity.

- Condition Number < 30 → No serious multicollinearity.
- Condition Number > 30 → possible multicollinearity.
- Condition Number > 100 → severe multicollinearity.

Scatterplots:

A scatter plot (aka scatter chart, scatter graph) uses dots to represent values for two different numeric variables. The position of each dot on the horizontal and vertical axis indicates values for an individual data point. Scatter plots are used to observe relationships between variables. Plotting pairs of independent variables against each other to visually identify linear relationships that may indicate multicollinearity.

Condition Index:

Condition index can refer to a measure of multicollinearity in a regression model, a way to assess the health of an individual or asset, or a tool for assessing the condition of a facility. A more advanced method that provides information about the severity of multicollinearity by assessing how close the correlation matrix is to being singular.

- CI above 30 indicates strong multicollinearity
- CI between 10 and 30 indicates low to moderate multicollinearity.

Ridge Regression:

Ridge regression reduces multicollinearity by adding a penalty term to the model. Does not detect collinearity directly but mitigates its effects.

- If Ridge regression shrinks coefficients significantly, multicollinearity is present.

Principal Component Analysis (PCA)

PCA can transform correlated variables into uncorrelated principal components. While PCA removes collinearity, it makes interpretation harder.

- If one or two components explain most of the variance, multicollinearity exists.
- PCA helps in dimensionality reduction to remove redundant features.

Key Takeaways:

1. Start with a correlation matrix to check for high correlations.
2. Use VIF to identify problematic variables.
3. Apply Ridge regression or PCA to fix multicollinearity if needed.