MULTICOLLINEARITY

Overview

Multicollinearity occurs when independent variables in a dataset are highly correlated. This can distort the importance of individual predictors in a regression model and lead to unreliable statistical inferences.

Detecting Multicollinearity

1. Correlation Matrix

Use a correlation matrix to visually inspect relationships between features. Check for high pairwise correlations (e.g., above 0.8 or 0.9).

2. Variance Inflation Factor (VIF)

Quantifies how much a variable is inflated due to multicollinearity.

· VIF > 5 or 10 indicates a high level of multicollinearity.

Formula

$$VIFi = 1/1-Ri2$$

Techniques to Handle Multicollinearity

1. Drop Highly Correlated Features

If two features are highly correlated (e.g., ssc_p and hsc_p), consider removing one.

2. Principal Component Analysis (PCA)

PCA transforms the original variables into a new set of **uncorrelated variables** called **principal components**.

These components:

- Are linear combinations of the original variables
- Are ordered by the amount of variance they capture (PC1 captures the most)
- Help simplify data while retaining most of its important information

3. Ridge Regression

Applies L2 regularization to reduce coefficient instability.Penalizes large coefficients and reduces variance.

4. Lasso regression: Can shrink some coefficients to zero, effectively selecting variables.

5. Combine Correlated Features

Create a new feature from multiple correlated features.

6. Partial Least Squares (PLS)

Reduces multicollinearity while predicting target variable.